Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Análise do desempenho dos alunos da UNICAMP do vestibular à conclusão do curso utilizando U-Estatísticas.

Rafael Pimentel Maia
Orientadora: Profa. Dra. Hildete Prisco Pinheiro

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

Campinas - SP
2008

Desempenho dos alunos da UNICAMP do ingresso à conclusão do curso utilizando U-Estatísticas

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Rafael Pimentel Maia e aprovada pela comissão julgadora.

Campinas, 27 de Março de 2008.

Hildete Pinheiro

Profa. Dra. Hildete Prisco Pinheiro

Orientadora

Banca examinadora:

1. Profa. Dra. Hildete Prisco Pinheiro (Orientadora) - IMECC/UNICAMP
2. Prof. Dr. Dalton Francisco de Andrade - CTC/UFSC
3. Prof. Dr. Renato Hyuda Luna Pedrosa - IMECC/UNICAMP

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Júlia Milani Rodrigues

Maia, Rafael Pimentel
M28a Análise do desempenho dos alunos da UNICAMP do ingresso à conclusão do curso utilizando U-Estatística / Rafael Pimentel Maia -- Campinas, [S.P. :s.n.], 2008.

Orientador : Hildete Prisco Pinheiro
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Ação afirmativa. 2. Desempenho acadêmico. 3. Medidas de diversidade. 4. Estatística não paramétrica. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Analysis of the students performance at UNICAMP from entrance to conclusion using U-Statistics.

Palavras-chave em inglês (Keywords): 1. Affirmative action. 2. Academic performance. 3. Diversity measures. 4. Nonparametric statistics.

Área de concentração: Probabilidade e Estatística Aplicada

Titulação: Mestre em Estatística

Banca examinadora:

Profa. Dra. Hildete Prisco Pinheiro (IMECC/UNICAMP)
Prof. Dr. Dalton Francisco de Andrade (CTC/UFSC)
Prof. Dr. Renato Hyuda Luna Pedrosa (IMECC/UNICAMP)

Data da defesa: 27/03/2008

Programa de pós-graduação: Mestrado em Estatística

Dissertação de Mestrado defendida em 27 de março de 2008 e aprovada pela Banca Examinadora composta pelos Profs. Drs.

Hildete Pinheiro

Prof (a). Dr (a). HILDETE PRISCO PINHEIRO

Prof (a). Dr (a). DALTON FRANCISCO DE ANDRADE

Prof (a). Dr (a). RENATO HYUDA DE LUNA PEDROSA

Aos meus pais José e Josefina e a minha esposa Érika.

Agradecimentos

A Deus por ter me concedido mais essa conquista.

Aos meus pais, José e Josefina, pelo apoio e incentivo que sempre me deram.

À minha querida esposa Érika, pelo apoio e paciência que teve comigo, principalmente no processo de conclusão desta dissertação.

À cara professora Hildete, pela orientação, pela paciência e, principalmente, pela amizade.

Ao professor Noberto Dachs, pessoa por quem tenho grande admiração, pela motivação e pela amizade.

Ao professor Aluisio Pinheiro pela ajuda no desenvolvimento da metodologia.

Ao professor Johan René Van Dorp da Universidade de George Washington, pela ajuda em adequar o software *MLE Calculator* aos dados do estudo.

Aos Professores Dalton Andrade e Renato Pedrosa, pelas críticas e sugestões que foram recebidas com muito respeito.

A todos os familiares e amigos que de alguma forma contribuiram para a conclusão desse trabalho. Principalmente aos que tiveram paciência de ouvir minhas lamentações!!

À Capes pelo suporte financeiro.

Muito Obrigado

"As nossas dúvidas são traidoras e nos fazem perder o bem que poderiamos conquistar, se não fosse o medo de tentar."

William Shakspeare

Resumo

O objetivo deste trabalho é propor novas metodologias para avaliar o desempenho dos alunos da UNICAMP, do ingresso à conclusão do curso. O conjunto de dados disponível foi obtido a partir dos questionários Sócio-Culturais aplicados pela Comissão Permanente de Vestibulares (COMVEST) na inscrição do vestibular e informações acadêmicas fornecidas pela Diretoria Acadêmica (DAC) da UNICAMP. Estes se referem às informações de todos os alunos ingressantes nos anos de 1997 a 2000.

São propostas duas metodologias, uma com base na variável denominada "ganho relativo" sugerido por Dachs e Maia (2006) e a segunda utilizando as notas de todas as disciplinas cursadas pelos alunos durante a graduação. Essas novas metodologias baseiam-se em medidas de diversidades propostas por Rao (1982) e na utilização de U-Estatísticas. São propostos testes de homogeneidade para avaliar se existe diferença no desempenho entre alunos de grupos distintos (alunos oriundos de escola pública ou privada, por exemplo). Aspectos teóricos de U-Estatística e medidas de diversidade também são apresentados.

Para a primeira metodologia foram feitas duas abordagens: paramétrica e não paramétrica, enquanto que para a segunda, apenas a abordagem não paramétrica foi explorada. Na abordagem paramétrica as estimativas são feitas por máxima verossimilhança e na não paramétrica foi utilizado o método de re-amostragem por ***jackknafe*** para se obter as estimativas das variâncias. Todas as aplicações utilizaram os dados dos alunos ingressantes.

Palavras-Chave: Ação Afirmativa, Desempenho Acadêmico, Medidas de Diversidade e Estatísticas Não Paramétrica.

Abstract

The main interest of this work is to propose new methods to evaluate the preformances of the students at UNICAMP from admission to graduation. The data was obtained from questionaires applied by the University Commission of admission's exam (COMVEST) during registration of the exam and academic informations provided by the Directory of Academic Studies (DAC). The data refer to information with respect to all the students enrolled in the University from 1997 to 2000.

We propose two methods: one based on the variable "relative gain"(Dachs and Maia, 2006) and the other method uses information about the grades of all courses attended by the students during their undergraduate studies. These new methods are based on diversity measures proposed by Rao (1982) and the use of U-Statistics. Homogeneity tests are proposed to evaluate differences in the performance of the students according to different socio-economic groups.

For the first method, we have two approaches: a parametric and a nonparametric analysis. For the second method, only a nonparametric analysis was done. In the parametric analysis, a Maximum Likelihood Estimation procedure is used and in the nonparametric analysis, resampling methods such as *jackknife* was used to obtain the estimates of the variances and confidence intervals. All the applications use the data of the enrolled students.

Keywords: Affirmativa action, Academic performance, Diversity measures and Nonparametric statistics.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Na sociedade brasileira tem se constatado há alguns anos muitas discussões sobre medidas para se reparar algumas das injustiças sociais históricas em relação ao Ensino Superior no país. As principais propostas tem sido no sentido de estabelecer "quotas". A médio e longo prazo, as políticas públicas para reverter este quadro devem concentrar-se em diminuir as desigualdades sociais existentes na sociedade e, neste caso em particular, em aumentar o acesso ao ensino médio, melhorando a qualidade do mesmo nas escolas públicas do país. Com esse objetivo, diversas medidas vem sendo tomadas a fim de que no futuro jovens brasileiros de ambos os sexos, que não têm a oportunidade de estudar em escolas privadas, possam ter as mesmas oportunidades de acesso à Universidade que os mais privilegiados economicamente. A mesma discussão ocorre também em relação ao acesso de pessoas que se auto-declaram negras ou pardas.

Em algumas universidades no país adotou-se a política de "quotas" para estudantes oriundos de escolas públicas e/ou auto-declarados negros. Na UNICAMP, a partir de 2004, foi tomada a decisão de, em vez de quotas, adotar políticas do tipo ação afirmativa (Bowen e Bok, 1998). Foi então criado um programa chamado **PAAIS** (Programa de Ação Afirmativa e Inclusão Social), adicionando um determinado número de pontos à nota do vestibular (esses pontos são adicionados a nota final, após a segunda fase) para aqueles candidatos que tenham cursado o ensino médio integralmente em escolas da rede

pública de ensino.

Dachs e Maia (2006) propuseram modelos de regressão com o objetivo de fornecer maiores subsídios sobre a adequação deste tipo de política. Para isso foi utilizado um conjunto de dados fornecido pela COMVEST (Comissão Permanente para os Vestibulares), com informações de todos os alunos ingressantes na universidade nos anos de 1994 a 1997. A variável de interesse, denominada de desempenho ou "ganho relativo", foi construída a partir da diferença do posto relativo do aluno referente ao seu coeficiente de rendimento(CR) final (razão entre a colocação do aluno e o número de alunos da turma), dentro de sua turma, e o posto relativo do aluno referente à nota final padronizada (NFP) obtida no vestibular.

O objetivo deste trabalho é propor novas metodologias para avaliar as diferenças com relação ao desempenho dos alunos. Essas novas metodologias se baseiam na utilização de *medidas de diversidade* ou *Análise de diversidade* (Rao, 1982). As medidas de diversidade têm sido muito utilizadas em diversas áreas do conhecimento (Pinheiro e Pinheiro, 2007), tais como: antropologia (Mahalanobis, 1936), genética (Cavalli-Sforza, 1969 ; Karlin et all, 1979 ; Nei, 1972), economia (Gini, 1912 ; Nayak e Gastwirth, 1989; Sen, 1973; Sen, 1999), sociologia(Agresti e Agresti, 1978 ; Rao, 1982) e outras áreas da biologia (Pinheiro, 1997 ; Shangvi, 1953; Sokal e Snealth; 1963). A análise da diversidade pode ser considerada como uma generalização da análise clássica de variância (ANOVA) e vem sendo muito útil para se analisar dados qualitativos que surgem nessas diversas áreas.

Foram propostas então duas metodologias, uma utilizando o "ganho relativo" e outra as notas obtidas pelos alunos nas disciplinas cursadas durante a graduação e o posto do aluno com relação a nota do vestibular. Ambas serão melhor apresentadas em capítulos seguintes.

A partir da análise de diversidade pretende-se avaliar se existem diferenças com relação ao desempenho acadêmico, entre alunos que estudaram o segundo grau em escolas públicas e escolas particulares.

Os dados utilizados nesse estudo foram fornecidos pela Comissão Permanente de Ves-

tibulares (COMVEST) e pela Diretoria Acadêmica (DAC) e se referem a informações acadêmicas e as respostas do questionário sócio-cultural aplicado no momento da inscrição do vestibular, de todos os alunos ingressantes na UNICAMP nos anos de 1997, 1998, 1999 e 2000. Esses dados correspondem à uma amostra do universo de todos os alunos que já ingressaram na UNICAMP.

1.1 Organização do Trabalho

O trabalho está dividido em 5 capítulos. O primeiro capítudo apresenta uma introdução como motivação para o trabalho, descrição da variável *ganho relativo* e um estudo de sua distribuição.

No segundo capítulo são apresentados alguns pontos importantes da teoria de U-Estatística, como a definição de uma U-Estatística, o teorema da decomposição de Hoeffding (Hoeffding, 1948) e alguns teoremas que garantem a convergência assintótica para uma distribuição nomal.

No capítulo 3 se discute um pouco dos aspectos teóricos da utilização de medidas de diversidades e da análise de diversidade, como uma alternativa não paramétrica.

No capítulo 6 são expostas, em detalhes, as duas metodologias propostas para a análise. E descritas as abordagens paramétricas e não paramétricas para se estimar a variância dos estimadores.

As aplicações são apresentadas no capítulo 5, onde é feito uma descrição detalhada do conjunto de dados e são aplicadas, a partir de dados reais, as metodologias propostas.

Nos apêndices é apresentada a Distribuição Triangular, além das rotinas desenvolvidas no software R para a aplicação das metodologia propostas.

1.2 O *"ganho relativo"*

Ao se pensar no problema de avaliar o desempenho dos alunos da UNICAMP, Dachs e Maia (2006) precisavam criar uma quantidade que mensurasse, de forma comparável entre os diferentes cursos e anos de ingresso, tal desempenho. As variáveis que dispunham eram a *nota final do vestibular* e o *coeficiente de rendimento* (CR) do aluno.

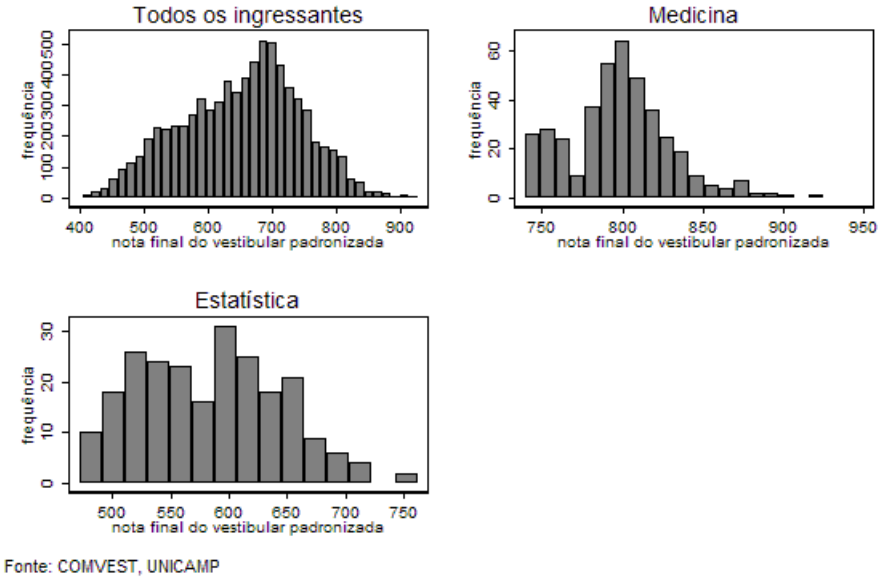

Figura 1.1: Histogramas para as notas finais padronizadas de todos os ingressantes dos anos de 1997, 1998, 1999 e 2000, e alunos aprovados nos Cursos de Medicina e Estatística, respectivamente.

A nota final padronizada do vestibular é a média das notas nas provas das diferentes disciplinas, padronizada para ter média 500 e desvio padrão 100. Como se observa na Figura 1.1, esta variável tem para o conjunto de todos os alunos que ingressaram na universidade uma distribuição, que é o resultado de uma mistura complexa, que pode possivelmente ser aproximada por uma mistura de Normais truncadas, com pontos de truncamento e médias diferentes para cada Curso. Mas descobrir essa forma não resolve o problema da comparabilidade. O mais importante é que para poder comparar ingressantes de Cursos diferentes não se pode usar a própria nota. Além disso a nota obtida no

vestibular avalia o aluno apenas no ingresso à universidade.

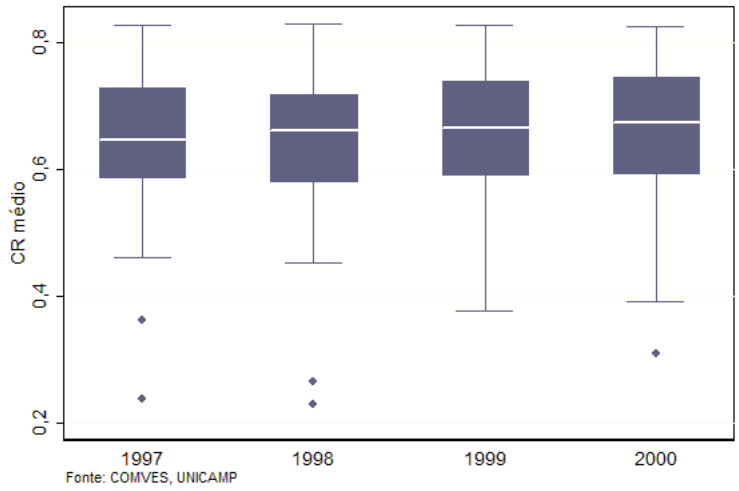

Figura 1.2: Gráficos de caixa para o CR médio dos alunos segundo o ano de ingresso na UNICAMP.

Uma situação ainda mais complexa ocorre com o coeficiente final de rendimento (CR) do aluno. Os processos de avaliação interna são muito diferentes para os Cursos das várias áreas e também apresentam variações ao longo do tempo (ver Figura 1.2). Essa característica pode ser observada a partir do CR médio dos alunos por turma. Os valores oscilam desde um mínimo próximo de 0,4 (Física e Matemática Licenciatura, Engenharia Agrícola e Estatística, dependendo do ano) até um máximo ao redor de 0,8 (Medicina, Enfermagem e Pedagogia diurno, dependendo do ano).

Por estas razões, foi decidido criar uma variável que foi chamada de *desempenho relativo* ou *ganho relativo*, da seguinte forma: aos alunos de uma mesma turma (ingressantes em mesmo ano e curso) foram atribuídos dois postos (colocação), um baseado na nota final do vestibular e outro no CR (o aluno com menor nota recebeu o posto 1, segunda menor nota o posto 2, e assim sucessivamente), chamados de *posto inicial e final*, respectivamente. Os postos foram divididos pelo número total de alunos em cada turma, para torná-los comparáveis entre turmas, já que as mesmas variam de tamanho de um curso e ano para o outro, padronizando a escala dos postos para todas as turma, fazendo

estes variarem entre 0 e 1. Esses novos valores foram chamados de **postos relativos**. Desta forma, o aluno com maior CR ou nota do vestibular recebe posto relativo igual a 1 e, quanto mais próximo de 1 for o posto relativo (inicial ou final) do o aluno, melhor foi o seu desempenho com relação a sua turma, e quanto mais próximo de 0, pior o seu desempenho.

O *ganho relativo* foi obtido da diferença entre o posto relativo baseado no CR (*posto relativo final*) e o posto relativo baseado na nota final do vestibular (*posto relativo inicial*). O ganho relativo é, por construção, uma variável limitada entre -1 e 1 e simétrica em torno do *zero*, com média, mediana e moda iguais a zero. Apesar do problema de que os cursos têm métodos de avaliações distintos e as turmas possuem diferentes tamanhos (número de alunos), é razovel assumir que a variável *ganho relativo*, da maneira como foi construída, seja comparável entre turmas. Isso porque se trabalha com postos (ao invés das notas absolutas) relativos a sua turma, ou seja, padronizados para variarem entre 0 e 1.

O conjunto de dados fornecido pela COMVEST e a DAC da UNICAMP, contém informação de 7515 alunos ingressantes nos anos de 1997 a 2000 (foram excluídos da amostra os alunos ingressantes em cursos tecnológicos pertencentes ao Campos de Limeira). Algumas medidas resumo da variável ganho relativo criado a partir desses dados, são apresentadas na Tabela 1.1.

A Figura 1.3 mostra os histogramas para o posto relativo inicial e final. Como se observa, as duas distribuições são idênticas (pelo método de construção das mesmas) e se aproximam de uma distribuição Uniforme(0,1).

Tabela 1.1: Medidas resumo para o ganho relativo da amostra total.

Variável	**n**	**mediana**	**média**	**desvio padrão**	**mínimo**	**máximo**
ganho relativo	7515	0,0000	0,0000	0,3543	-0,9843	0,9839

Na Figura 1.4 é apresentado o histograma para o ganho relativo e o gráfico *"Quantil-Quantil da normal"* do mesmo. Nota-se que o ganho relativo têm uma distribuição com as caudas mais leves comparadas a uma distribuição normal, o que era de se esperar por

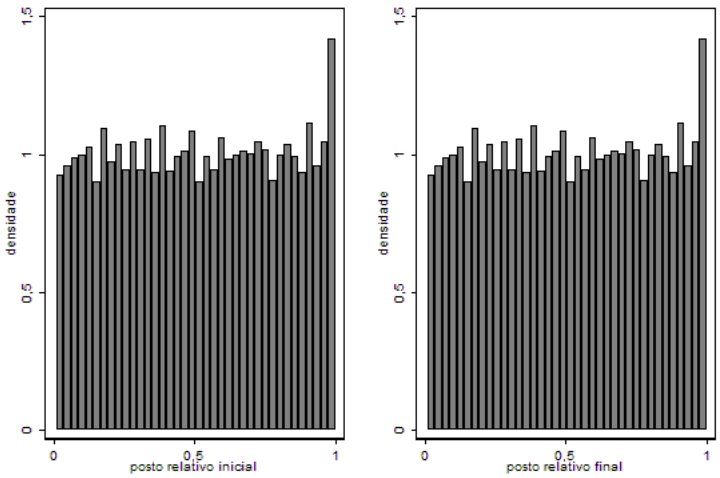

Figura 1.3: Histogramas para o posto relativo obtido a partir da nota final do vestibular padronizada e do coeficiente de rendimento, respectivamente.

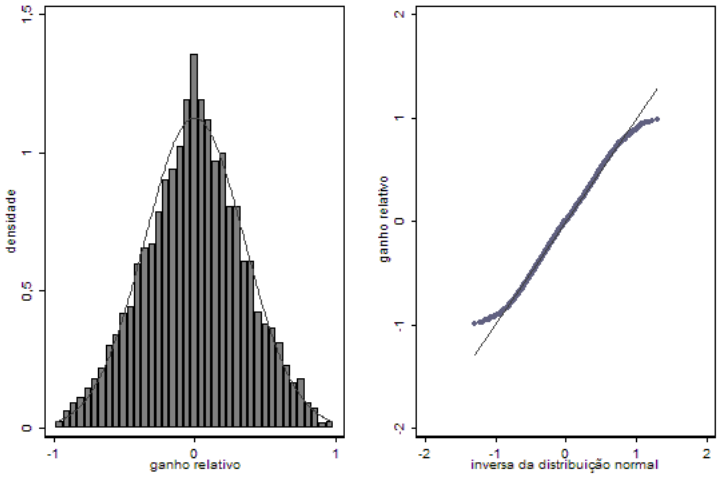

Figura 1.4: Histograma e *Gráfico Quantil-Quantil normal* para o ganho relativo.

tratar-se de uma variável limitada a um intervalo. Como a distribuição do ganho relativo é limitada e unimodal, uma suposição razoável é assumir que vem de uma distribuição triangular em [-1,1] com moda igual a *zero* (Kotz e Dorp, 2004). Um outro fator que justifica tal suposição, é que a distribuição triangular pode ser obtida a partir da subtração de duas variáveis com distribuições Uniformes, e as distribuições dos postos relativos se aproximam de Uniformes em (0,1).

A Figura 1.5 mostra o histograma para ganho relativo com curva de uma densidade triangular[-1,0,1] e o gráfico *Quantil-Quantil da triangular[-1,0,1]* do mesmo. O desvio padrão de uma variável com distribuição triangular em [-1,0,1] é igual a 0,4082 que é maior que o desvio padrão observado na amostra (0,3543).

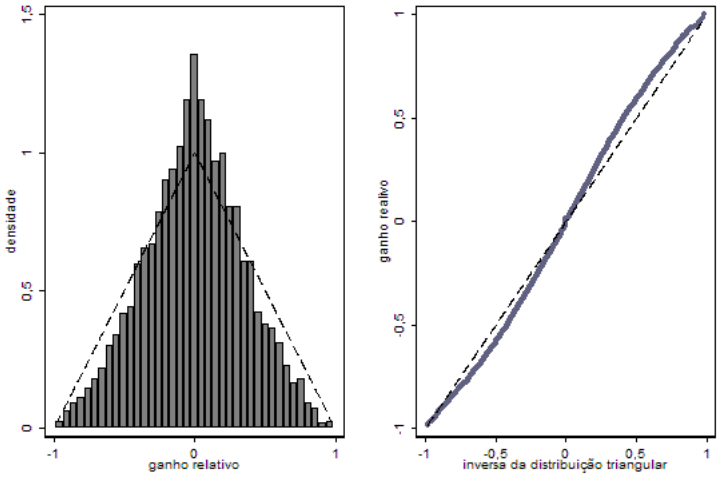

Figura 1.5: Histograma e *Gráfico Quantil-Quantil triangular[-1,0,1]* para o ganho relativo.

Para averiguar se de fato o ganho relativo vem de uma distribuição triangular, foi gerada uma amostra aleatória da seguinte forma: para cada aluno dentro de uma mesma turma foi dado uma posição inicial e uma final totalmente aleatória (como em um sorteio), desta maneira os postos (iniciais e finais) têm uma distribuição uniforme discreta. Nos dados o número total de alunos é de 7515, divididos em 183 turmas distintas (em 4 anos de ingresso: 1997, 1998, 1999 e 2000), sendo assim a amostra gerada também tem $\mathbf{n} = 7515$.. Em seguida essas posições foram divididas pelo número de alunos em cada turma. O

ganho relativo "gerado" foi então obtido da subtração entre os dois postos.

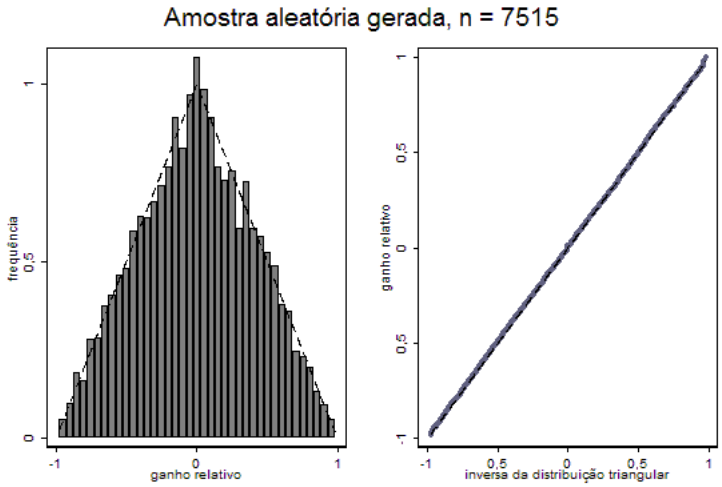

Figura 1.6: Histograma e *Gráfico Quantil-Quantil triangular[-1,0,1]* para a amostra aleatória do ganho relativo.

Como se nota na Figura 1.6 a amostra aleatória gerada segue perfeitamente uma distribuição triangular, o que não ocorre na figura anterior com os dados reais. O que se observa é que nos dados a percentagem de alunos que tiveram ganho igual a 0 (não mudaram de posição) ou ganho próximo de zero (mudaram poucas posições) é maior do que o esperado em uma amostra onde os postos são atribuídos de forma totalmente aleatória.

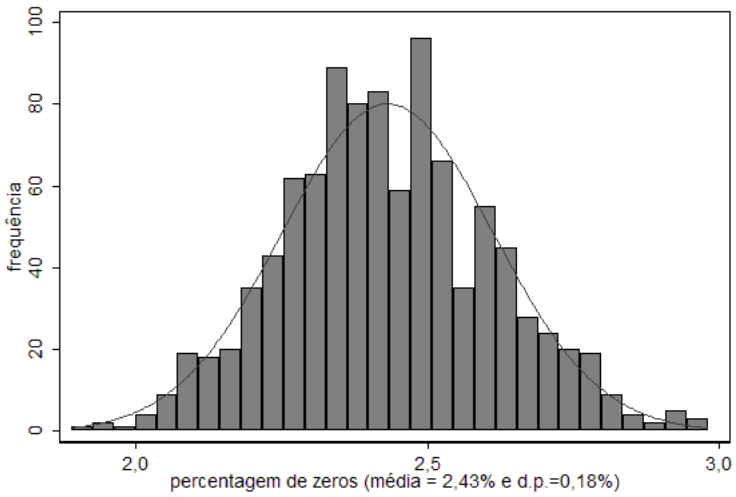

Figura 1.7: Histograma da percentagem de zeros nas 1000 amostras geradas do ganho relativo.

A partir dos resultados observados com respeito a suposição de que o ganho relativo têm distribuição triangular foi avaliado apenas os alunos que não tiveram alteração de postos (ganho relativo igual a 0) - nos dados eles correspondem a 3,65% da amostra. Foram então geradas 1000 novas amostras (utilizando o procedimento anterior) aleatórias do ganho relativo de tamanho 7515 e registrada a percentagem de alunos com ganho igual a zero em cada uma delas.

O histograma dessas percentagens é apresentado na Figura 1.7. A média é igual a 2,43% (desvio padrão = 0,18%). Aplicando o teste de normalidade de Sahpiro-Wilk o p-valor obtido é 0,1280, portanto não se rejeita a hipótese de normalidade, ao nível de 5%. Em seguida foi aplicado o teste t-Sudent para avaliar a probabilidade da média da amostra de percentagem de zeros ser igual a 3,65% e o p-valor encontrado foi $< 0{,}0001$. Portanto, existem evidências para se rejeitar a hipótese nula de que a média da percentagem de zeros é igual a 3,65%.

Baseando-se nessa informação, outra sugestão foi aplicar uma correção na função de distribuição do ganho, da seguinte forma: seja $F(\cdot)$ a função de distribuição triangular em [-1,1] com *moda* igual 0, e seja p a probabilidade do ganho relativo ser igual a zero. A função de distribuição ajustada para o ganho relativo seria, então, dada por

$$F^*(x) = \begin{cases} 0 & se \quad x < -1 \\ (1-p)F(x) & se \quad -1 \leq x < 0 \\ p + (1-p)F(x) & se \quad 0 \leq x \leq 1 \\ 1 & se \quad x > 1 \end{cases} \tag{1.1}$$

onde o valor estimado para p foi 0,03646 (3,65%).

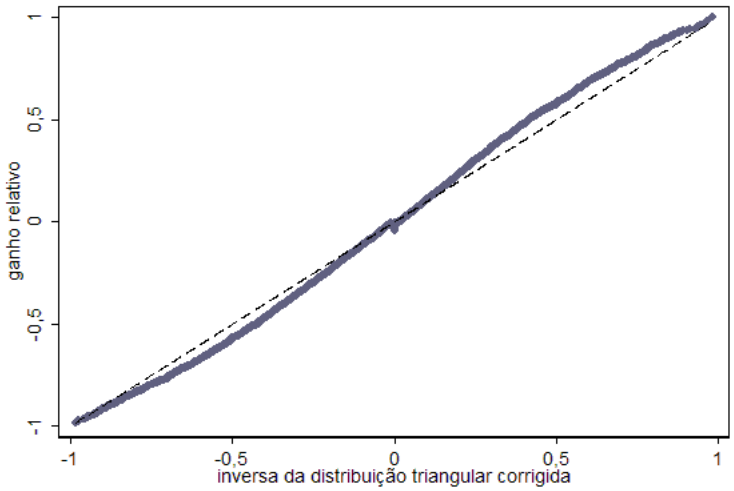

Figura 1.8: *Gráfico Quantil-Quantil triangular[-1,0,1] corrigida*, para o ganho relativo.

Mesmo fazendo esta correção, como se observa na Figura 1.8, ainda há problemas no ajuste da distribuição do ganho, isso porque as caudas da distribuição não tem um decaimento linear, como era esperado que tivesse.

Essa análise mostra que a suposição de que o ganho relativo vêm de uma distribuição triangular em [-1,0,1] é razoável, entretanto nos dados presentes a percentagem de alunos com ganho muito pequeno é maior do que o esperado, o que distorce a distribuição real. Desta forma a distribuição normal parece melhor se adequar aos dados. No entanto, irá se trabalhar também com a distribuição triangular na abordagem paramétrica feita para o ganho relativo.

Após cuidadosa análise da distribuição do dados, será explorado também uma abordagem não paramétrica, onde será utilizada a técnica de re-amostragem de ***jackknife*** para

se obter as estimativas das variâncias dos estimadores da estatística do teste que será proposta, e a partir da variância estimada calcular os respectivos intervalos de confiança.

Capítulo 2

U-Estatísticas

A idéia básica que norteia a classe de U-Estatísticas, é a representação de uma característica populacional de interesse como funcional da função de distribuição (Pinheiro e Pinheiro, 2007).

Suponha um conjunto $\mathcal{F}$ de funções de distribuição. Defina-se um funcional $\theta(\cdot)$ em $\mathcal{F}$ por

$$\theta = \theta(F), \quad F \in \mathcal{F}$$

Halmos (1946) demonstra o seguinte teorema. Seja $\mathcal{F}$ um conjunto qualquer de funções de distribuição em $\mathbb{R}$. Considere θ um funcional definido em $\mathcal{F}$. Seja ainda $X_1, ..., X_n$ uma amostra aleatória de F. Então,

Teorema 2.1 (Halmos,1946)**.**

Um funcional θ definido em $\mathcal{F}$ pode ser estimado sem vício se e somente se existe uma função ϕ (em $\mathbb{R}^k$, para algum k) tal que

$$\theta(F) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \phi(x_1, ..., x_k) dF(x_1) ... dF(x_k),$$

para todo $F \in \mathcal{F}$.

Um funcional satisfazendo o Teorema 2.1 é chamado de um funcional estatístico *regular* de grau k. A função ϕ associada é dita *núcleo* do funcional. Isto leva à seguinte definição de *parâmetros estimáveis*.

Definição 2.1 (Parâmetros Estimáveis - Pinheiro e Pinheiro, 2007)**.**

Sejam $\theta \in \mathbb{R}$ e $X_1, X_2, ...$ uma seqüência de variáveis aleatórias independentes e identicamente distribuídas F, para algum F na família de distribuições $\mathcal{F}$. Então, diz-se ser θ um parâmetro estimável de grau r na família de distribuições $\mathcal{F}$ se existe um núcleo $\phi(\cdot, ..., \cdot)$ (de r argumentos) tal que

$$E_F(\phi(X_1, ..., X_r)) = \theta,$$

para todo $F \in \mathcal{F}$.

2.1 Definindo U-Estatísticas

Sejam $X_i, i \geq 1$, variáveis aleatórias independentes e identicamente distribuídas com função de distribuição F, definida no $\mathbb{R}^p$, para algum $p \geq 1$. Considere a função $\theta(F) = \theta$ definida por

$$\theta(F) = \int ... \int \phi(x_1, ..., x_m) dF(x_1) ... dF(x_m),$$

em que $\phi(x_1, ..., x_m)$ é uma função simétrica de $m \leq 1$ argumentos.

Por exemplo, se $\phi(x) = x$, isto é, $m = 1$, então $\theta(F) = EX_1 = \mu$. Similarmente, se $m = 2$ e $\phi(x_1, x_2) = \frac{(x_1 - x_2)^2}{2}$ então, $\theta(F) = E(X_1 - X_2)^2/2 = E[(X_1 - \mu) - (X_2 - \mu)]^2/2$ $= E(X - \mu)^2 = \sigma^2$.

Desta forma, grande parte dos parâmetros podem ser formulados como funções de suas funções de distribuições latentes. Note que é equivalente a dizer que,

$$\theta(F) = E_F(\phi(X_1, ..., X_m)),$$

para todo F pertencente a classe $\mathcal{F}$.

Nesta forma, $\phi(\cdot)$ é chamada de função núcleo (ou *kernel*) de grau m. Se $X_1, ..., X_n$ é uma amostra de tamanho n, e toma-se alguma sub-amostra m ($n \geq m$), estima-se $\theta(F)$ por $\phi(X_{i_1}, ..., X_{i_m})$. Portanto, um estimador simétrico e não viciado de $\theta(F)$ pode ser obtido por combinar todos estes estimadores não viciados. Isto é chamado de *U-Estatística* e é dado pela seguinte definição,

Definição 2.2 (U-Estatística).

$$U^m \equiv U(X_1, ..., X_m) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < ... < i_m \leq m} \phi(X_{i_1}, ..., X_{i_m}), \quad n \geq m.$$

Exemplos

(Média Amostral). Se $\theta(F) = \mu$ e $\phi(X) = X$, então

$$U^1 = \binom{n}{1}^{-1} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n.$$

(Variância Amostral). Se $\theta(F) = \sigma^2$ e $\phi(X_1, X_2) = \frac{(X_1 - X_2)^2}{2}$, então

$$\begin{aligned}
U^2 &= \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(X_i - X_j)^2}{2} \\
&= \frac{1}{n(n-1)} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_i^2 - 2X_iX_j + X_j^2 \right] \\
&= \frac{1}{n(n-1)} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2X_iX_j + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_j^2 \right] \\
&= \frac{1}{n(n-1)} \left[\sum_{i=1}^{n-1} X_i^2(n-i) + \sum_{j=2}^{n} X_j^2(i-1) - \left(\sum_{i=1}^{n} \sum_{j=1}^{n} X_iX_j - \sum_{i=1}^{n} X_i^2 \right) \right] \\
&= \frac{1}{n(n-1)} \left[n \sum_{i=1}^{n} X_i^2 - n^2 \bar{X}^2 \right] \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} \\
&= \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \\
&= S_n^2.
\end{aligned}$$

Definição 2.3 (U-Estatística generalizada).

Considere k amostras aleatórias independentes, $\{X_1^1, X_2^1, ..., X_1^k, X_2^k, ...\}$*, obtidas das distribuições* $F_1, ..., F_k$*, respectivamente. Seja a função paramétrica* $\theta = \theta(F_1, ..., F_k)$*, da qual existe um estimador não viciado. Isto é,*

$$\theta = E[\phi(X_1^1, ..., X_{m_1}^1; ...; X_1^k, ..., X_{m_k}^k)],$$

em que, sem perda de generalidade, ϕ é simétrico dentro de cada um dos k grupos de argumentos. Para esse núcleo ϕ, assumindo que $n_1 \geq m_1, ..., n_k \geq m_k$, a U-estatística para θ é definida por

$$U^{\boldsymbol{m}} = \frac{1}{\prod_{j=1}^{k} \binom{n_j}{m_j}} \sum_{c} \phi(X_{i_1 1}^{1}, ..., X_{i_1 m_1}^{1}; ...; X_{i_k 1}^{k}, ..., X_{i_k m_k}^{k}),$$

em que $\boldsymbol{m} = \{m_1, m_2, ..., m_k\}$ e $\{i_{j1}, ..., i_{jm_i}\}$ denota um conjunto de m_j elementos distintos do conjunto $\{1, 2, ..., n_j\}$, $1 \leq j \leq k$ e $\sum_c$ denota a soma sobre todas as combinações.

Exemplo. U-Estatística generalizada de grau (1,1).

Estatística de Wilcoxon para 2 grupos. Seja $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ amostras aleatórias das distribuições F_1 e G_1, respectivamente. Então o estimador não viciado de

$$\theta(F, G) = \int_{-\infty}^{\infty} \int_{x}^{\infty} dF dG = P(X \leq Y)$$

é

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n1} \sum_{j=1}^{n_2} I(X_i \geq Y_j).$$

2.2 Propriedades de U-Estatísticas

Defina-se as funções

$$\phi_c = E[\phi(X_1, ..., X_k) | X_1 = x_1, ..., X_c = x_c] = E[\phi(x_1, ..., x_c, X_{c+1}, ..., X_m)],$$

para $c = 1, 2, ..., m$

A Função ϕ_c apresenta as seguintes propriedades (Lee, 1990):

1. $\phi_c(x_1, ..., x_c) = E(\phi_d(x_1, ..., x_c, X_{c+1}, ..., X_d))), \ 1 \leq c < d \leq m$;
2. $E(\phi_c(x_1, ..., x_c)) = E(\phi(x_1, ..., x_c))$.

Defina-se agora

$$\sigma_0^2 = 0,$$

$$\sigma_c^2 = Var[\phi_c(X_1, ..., X_c)],$$

$c = 1, ..., m$

Um resultado interessante é o proposto por Hoeffding (1948), chamado de decomposição de Hoeffding ou, simplesmente, *Decomposição H*. Este resultado demonstra que toda U-Estatística de grau m pode ser escrita como uma combinação linear de U-Estatísticas não correlacionas de graus $1, 2, .., m$

Teorema 2.2 (Decomposição de Hoeffding)**.**

Seja $X_1, ..., X_n$ uma amostra aleatória da distribuição F e $\phi(\cdot)$ um núcleo de grau m. Definam-se

$$\psi_{(1)}(x_1) \;=\; \phi_1(x_1) - \theta,$$

$$\psi_{(2)}(x_1, x_2) \;=\; \phi_2(x_1, x_2) \;-\; \psi_{(1)}(x_1) \;-\; \psi_{(1)}(x_2) - \theta,$$

$$...$$

$$\psi_{(c)}(x_1, ..., x_c) \;=\; \phi_c(x_1, ..., x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} \psi_{(j)}(x_{i_1}, ..., x_{i_j}) - \theta,$$

para $c = 3, .., m$, em que $\sum_{(c,j)}$ é tomada para todos os subconjuntos de cardinalidade j de $\{x_1, ..., x_c\}$

A partir disso, pode-se escrever U como:

$$U = \theta + \sum_{j=1}^{k} \binom{k}{j} \Psi_{(j)},$$

onde $\Psi_{(j)} = \binom{n}{m}^{-1} \sum_{(n,j)} \psi_{(j)}(X_{i_1}, ..., X_{i_j})$.

Demonstração

Seja $S_{js_1,...,s_k} = \sum \psi_{(j)}(x_{i_1}, ..., x_{i_j})$, soma esta em todos os subconjuntos $\{i_1, ..., i_j\}$ de $\{s_1, .., s_k\}$. Note que,

$$\sum_{(n,m)} S_{js_1,...,s_m} \;=\; \binom{n-j}{m-j} \sum_{(n,j)} \psi_{(j)}(x_{i_1}, ..., x_{i_j}),$$

e, sendo verdadeira a relação

$$\binom{n}{m}^{-1}\binom{n-j}{m-j} = \binom{m}{j}\binom{n}{j}^{-1},$$

tem-se que

$$\begin{aligned} U &= \binom{n}{m}^{-1}\sum_{(n,m)}\phi(x_1,...,x_m) \\ &= \binom{n}{m}^{-1}\sum_{(n,m)}\Big(\sum_{j=1}^{m} S_{js_1,...,s_m} + \theta\Big) \\ &= \theta + \binom{n}{m}^{-1}\sum_{j=1}^{m}\binom{n-j}{m-j}\sum_{(n,j)}\phi_{(j)}(x_{i_1},...,x_{i_j}) \\ &= \theta + \sum_{j=1}^{k}\binom{k}{j}\Psi_{(j)}. \end{aligned} \tag{2.1}$$

□

2.3 Teoria Assintótica

Aqui serão apresentadas algumas definições da teoria assintótica utilizadas para obter os resultados assintóticos de U-Estatísticas apresentados na seção seguinte (Leite e Singer, 1990).

Definição 2.4 (Ordens de magnitude de seqüências de números reais e vetores)**.**

Sejam $\{a_n\}_{n\geq 1}$ *e* $\{b_n\}_{n\geq 1}$ *seqüências de números reais. Então diz-se que*

1. $a_n = O(b_n)$ *se existirem um número real* $K > 0$ *e um número inteiro positivo* $n_0 = n_0(K)$ *tal que* $|a_n/b_n| \leq K$, $\forall n \geq n_0$*;*

2. $a_n = o(b_n)$ *se para todo* $\epsilon > 0$ *existir um número inteiro positivo* $n_0 = n_0(\epsilon)$ *tal que* $|a_n/b_n| < \epsilon$, $\forall n > n_0$.

Definição 2.5 (Ordens de magnitude de seqüências estocásticas)**.**

Sejam $\{X_n\}_{n\geq 1}$ *uma seqüência de variáveis aleatórias e* $\{b_n\}_{n\geq 1}$ *uma seqüência de números reais (ou variáveis aleatórias). Diz-se que*

1. $X_n = O_p(b_n)$ *se para todo número real* $\eta > 0$ *existirem um número real positivo* $K = K(\eta)$ *e um número inteiro positivo* $n_0 = n_0(\eta)$, *tais que*

$$P(|X_n/b_n| \geq K) \leq \eta, \forall n \geq n_0;$$

2. $X_n = o_p(b_n)$ *se para todo número real* $\epsilon > 0$ *e para todo número real* $\eta > 0$ *existir um número inteiro positivo* $n_0 = n_0(\epsilon, \eta)$, *tal que*

$$P(|X_n/b_n| \geq \epsilon) < \eta, \forall n \geq n_0.$$

Teorema 2.3 (Variância de uma U-Estatística de grau m)**.**

Considere um núcleo de ordem m, $\phi(\cdot)$. *Então*

$$E\phi_c(X_1, ..., X_c) = \theta,$$

para todo $1 \leq c \leq m$

Mais ainda, sendo $\sigma_c^2 = Var(\phi_c(X_1, ..., X_c))$, *então*

1. *A variância de uma U-Estatística* U *pode ser escrita como*

$$\binom{n}{m} Var(U) = \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\sigma_c^2; \tag{2.2}$$

2. *Se* $\sigma_1^2 > 0$ *e* $\sigma_c^2 < \infty$, *para todo* $c = 1, ..., m$, *então*

$$Var(\sqrt{n}U) \to m^2\sigma_1^2, \text{ quando } n \to \infty. \tag{2.3}$$

Demonstração

Note que

$$\binom{n-m}{j} = \frac{1}{j!}(n-m)(n-m-1)...(n-m-j+1) \approx \frac{n^j}{j!}.$$

Portanto, na expressão 2.2, o termo principal corresponde a $c = 1$ que é assintoticamente equivalente a

$$\sigma_1^2 \frac{mn^{k-1}}{(m-1)!}\frac{m!}{n^m} = \frac{m\sigma_1^2}{n}.$$

□

Finalmente, o Teorema Central do Limite de Hoeffding é dado por

Teorema 2.4 (Normalidade Assintótica de U-Estatística)**.**

1. *Se* $0 < \sigma_1^2 < \infty$*, então, quando* $n \to \infty$

$$\sqrt{n}(U - \theta) \xrightarrow{D} N(0, m^2\sigma_1^2);$$

2. *Se* $\sigma_c^2 < \infty$*,* $c = 1, ..., m$*, então,*

$$\frac{U - \theta}{\sqrt{Var(U)}} \xrightarrow{D} N(0, 1).$$

Teorema 2.5 (Variância de U-Estatísticas Generalizadas)**.**

Considere uma U-Estatística generalizada $U^{\boldsymbol{m}}$*, faz-se a extensão da teoria assintótica para esse caso. Para isso, considere* d_j*, tal que* $0 \leq d_j \leq m_j.1 \leq j \leq k$*, se* $\boldsymbol{d} = (d_1, ..., d_k)$ *e*

$$\Phi_{d_1,...,d_k}(x_1^j, ..., x_{d_1}^j; 1 \leq j \leq k) = E(\phi(X_1^j, ..., X_{m_j}^j)|X_1^j = x_1^j, ..., X_{d_j}^j = x_{d_j}^j; 1 \leq j \leq k).$$

Portanto, $\Phi_0 = \theta(F)$*, pois* $\phi(X_1^j, ..., X_{m_j}^j; 1 \leq j \leq k)$ *é um estimador não viesado para* $\theta(F)$ *e* $\Phi_{\boldsymbol{m}} = \phi$*, com* $\boldsymbol{m} = (m_1, ..., m_k)$*. Então*

$$\varsigma_{\boldsymbol{d}} = E[\Phi_{\boldsymbol{d}}(X_1^j, ..., X_{d_j}^j; 1 \leq j \leq k)] - \theta^2(F), \quad \boldsymbol{0} \leq \boldsymbol{d} \leq \boldsymbol{m},$$

com $\varsigma_{\boldsymbol{0}} = 0$*. Portanto, para todo* $\boldsymbol{n} \leq \boldsymbol{m}$

$$Var(U^{\boldsymbol{m}}) = \sum_{j=1}^{k} n_j^{-1}\sigma_j^2[O(n_0^{-1})],$$

em que $n_0 = min(n_1, ..., n_k)$ *e* $\sigma_j^2 = m_j^2\varsigma_{\delta_{j1},...,\delta_{jk}}$*,* $j = 1, ..., k$ *com* $\delta_{\alpha,\beta} = 1$ *se* $\alpha = \beta$ *e* 0 *se* $\alpha \neq \beta$*.*

Então, se $E(\phi^2) < \infty$*,*

$$\gamma_{n_1,...,n_k}^{-1}(U^{\boldsymbol{m}} - \theta) \xrightarrow{D} N(0, 1).$$

quando $n_0 = min(n_1, ..., n_k) \to \infty$*, em que*

$$\gamma_{n_1,...,n_k}^2 = \sum_{j=1}^{k} \frac{m_1^2\varsigma_{\delta_{j1},...,\delta_{jk}}}{n_j}.$$

Com isso, a U-Estatística generalizada tem distribuição assintótica $N(\theta, \gamma_{n_1,...,n_k}^2)$*.*

Teorema 2.6 (Covariância entre duas U-Estatísticas de mesma amostra)**.**

Considere um conjunto de g U-Estatísticas,

$$U_\gamma = \binom{n}{m_\gamma} \sum_{(c)} \phi^\gamma(X_{\alpha 1}, ..., X_{\alpha m_\gamma}), \quad \gamma = 1, ..., g,$$

em que cada U_γ é função da mesma amostra aleatória de tamanho n $X_1, ..., X_n$. Assume-se que a função ϕ^γ é simétrica nos m_γ argumentos, $\gamma = 1, ..., g$. Sejam

$$E(U_\gamma) \;=\; E(\phi_\gamma(X_1, ..., X_{m_\gamma})) = \theta_\gamma, \quad \gamma = 1, ..., g;$$

$$\psi^\gamma(x_1, ..., x_{m_\gamma}) \;=\; \phi^\gamma(x_1, ..., x_{m_\gamma}) - \theta_\gamma;$$

$$\psi_c^\gamma(x_1, ..., x_{m_c}) \;=\; E[\psi\ \gamma(_1, ..., X_{m_\gamma})|X_1 = x_1, ..., X_c = x_c], \quad c = 1, ..., m_\gamma;$$

$$\varsigma_c^{\gamma,\upsilon} \;=\; E[\psi_c^\gamma(X_1, ..., X_c)\psi_c^\upsilon(X_1, ..., X_c)], \quad \gamma, \upsilon = 1, ..., g.$$

Em particular, se $\gamma = \upsilon$, então escreve-se,

$$\varsigma_c = \varsigma_c^{\gamma,\gamma} = E[\psi_c^\gamma]^2.$$

Seja,

$$\sigma(U_\gamma, U_\upsilon) = E[(U_\gamma - \theta_\gamma)(U_{\upsilon - \theta_\upsilon})],$$

a covariância entre U_γ e U_υ.

Se $m_\gamma < m_\upsilon$, da mesma forma que para a variância, encontra-se que,

$$\sigma(U_\gamma, U_\upsilon) \;=\; \binom{n}{m_\gamma}^{-1} \sum_{c=1}^{m_\gamma} \binom{m_\upsilon}{c} \binom{n - m_\upsilon}{m_\gamma - c} \varsigma_c^{\gamma,\upsilon}.$$

para $\gamma = \upsilon$, $\sigma(U_\gamma, U_\upsilon)$ é a variância de U_γ. Segundo Hoeffding (1948),

$$\lim_{n\to\infty} n\sigma(U_\gamma, U_\upsilon) = m_\gamma m_\upsilon \varsigma_1^{\gamma,\upsilon}.$$

Assim, pode-se fazer a seguinte aproximação:

$$\sigma(U_\gamma, U_\upsilon) \approx \frac{m_\gamma m_\upsilon}{n} \varsigma_1^{\gamma,\upsilon} + O(n^{-2}).$$

Capítulo 3

Medidas de Diversidade

Uma medida de diversidade pode ser usada para decompor a diversidade total dentro de uma determinada população devido a um certo número de fatores. Portanto pode-se perguntar quanto da diversidade entre indivíduos de uma população é devido ao tamanho e quanto é devido a forma.

No caso em que se tenha uma mistura de populações, pode-se estar interessado em saber quanto da diversidade da composição das populações é devido a diversidade dentro de cada população e quanto é devido a diversidade entre populações.

Em análise de variância divide-se a variabilidade em um dado conjunto de dados quantitativos dentro de um número de componentes aditivos, cada componente é usada para testar uma certa hipótese nula ou para estimar uma componente da variância. Rao (1982) introduziu uma medida geral de diversidade (variabilidade) aplicável tanto a dados quantitativos, quanto a dados qualitativos, estendendo o conceito de análise de variância (ANOVA) para um caso mais geral, chamando de análise de diversidade (ANODIV).

3.1 Aspectos Teóricos

Considere um espaço mensurável e um conjunto $\mathcal{P}$, convexo, de medidas de probabilidade definidas nele. Uma função $Q(\cdot)$ mapeando $\mathcal{P}$ nos reais é dita ser uma "medida de

diversidade" se esta satisfaz às seguintes condições

- C1: $Q(P) \geq 0 \ \forall \ P \in \mathcal{P}$ e $Q(P) = 0$ se, e somente se, P é degenerada;
- C2: Q é uma função côncava em $\mathcal{P}$.

$Q(P)$ será a diversidade dentro de uma população α caracterizada pela medida de probabilidade P. Considere agora uma função $\phi(X_1, X_2)$ simétrica e não negativa, que é uma medida de diferença entre dois indivíduos, sem dar referência a distribuição de probabilidade de X_1 e X_2. A escolha de $\phi(X_1, X_2)$ naturalmente depende da natureza do problema em questão. Rao (1982), define DIV (diversidade) da população i como

$$Q(P_i) \ = \ Q_i \ = \ \int\int \phi(x_1, x_2) dP_i(x_1) dP_i(x_2),$$

isto é, a diferença média entre dois indivíduos selecionados aleatoriamente da população i. Suponha que um indivíduo foi retirado da população i e o outro da população j. A diferença média entre esses dois indivíduos é dada por

$$C(P_i, P_j) \ = \ C_{ij} \ = \ \int\int \phi(x_i, x_j) dP_i(x_i) dP_j(x_j).$$

Espera-se que C_{ij} seja maior do que a média entre Q_i e Q_j, esse resultado é obtido a partir da *Desigualdade de Jensen* (se $\phi(x_i, x_j)$ é convexa)

$$C_{ij} \ \geq \ \frac{1}{2}(Q_i + Q_j).$$

A partir desse resultado, obtém-se a DIS (dissimilaridade) entre i e j, sendo definida como a *Diferença de Jensen*

$$D_{ij} \ = \ C_{ij} - \frac{1}{2}[Q_i + Q_j] \ \ ou \ \ 2D_{ij} = 2C_{ij} - [Q_i + Q_j].$$

A quantidade $D(\cdot, \cdot)$ será não negativa se $\phi(\cdot, \cdot)$ satisfizer algumas condições descritas a seguir.

Teorema 3.1 (Rao, 1984)**.**

Seja Q e D como descritas acima. Então

(a) D é não negativa para todo P_1 e P_2 se e somente se Q é uma função côncava no espaço das funções distribuições;

(b) Q é côncava se, e somente se, ϕ é uma função condicionalmente definida negativa (CDN), isto é, ϕ satisfaz a condição

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\phi(x_i,x_j)a_ia_j \geq 0$$

para todo $x_1,...,x_n$ e alguma escolha de números reais $a_1,...,a_n$, tal que $a_1+...+a_n=0$;

(c) ϕ É CDN se, e somente se, $\phi^{\frac{1}{2}}$ é uma métrica (i.e., satisfaz os axiomas de uma função distância); e

(d) se ϕ é CDN, então ϕ^{α} também é CDN para todo $0 \leq \alpha \leq 1$.

3.2 Estimadores de Q_i e C_{ij}

Seja $x_1, x_2, ..., x_m, y_1, y_2, ..., y_n$ amostras aleatórias de duas populações P_1 e P_2, então os estimadores para Q_1, Q_2, C_{12} e D_{12}, serão dados por

$$\hat{Q}_1 = \frac{1}{\binom{m}{2}}\sum_{i<j}\phi(x_i,x_j); \tag{3.1}$$

$$\hat{Q}_2 = \frac{1}{\binom{n}{2}}\sum_{i<j}\phi(y_i,y_j); \tag{3.2}$$

$$\hat{C}_{12} = \frac{1}{mn}\sum_{i}\sum_{j}\phi(x_i,y_j); \tag{3.3}$$

$$\hat{D}_{12} = \hat{C}_{12} - \frac{1}{2}(\hat{Q}_1+\hat{Q}_2). \tag{3.4}$$

$\hat{Q}_1$ e $\hat{Q}_2$ são U-Estatísticas de grau 2 e $\hat{C}_{12}$ é uma U-Estatística de grau (1,1).

3.3 No caso de 3 ou mais sub-populações

Sejam x_{11} , x_{12} , ... , x_{1n_1}, x_{21} , x_{22} , ... , x_{2n_2};; ;x_{R1} , x_{R2} , ... , x_{Rn_R}, R amostras aleatórias, obtidas de distribuições P_r, respectivamente, para $r = 1, ..., R$. Seja $n = \sum_{r=1}^{R} n_r$.

A Soma de Quadrados Total (SQT) ou *distância média geral* é a variabilidade total da amostra e pode ser escrita como

$$SQT \ = \ \sum_{i<j} \phi(x_i, x_j) \ = \ \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \phi(x_i, x_j). \tag{3.5}$$

A partir dos resultados apresentados na seção anterior, a distância média dentro do $r-$*ésimo* grupo é dada por

$$\hat{Q}_r = \binom{n_r}{2} \sum_{i=1}^{n_r-1} \sum_{j=i+1}^{n_r} \phi(x_i^r, x_j^r), \tag{3.6}$$

$$E(\hat{Q}_r) = Q(P_r) = \int \int \phi(x, y) dP_r(x) dP_r(y).$$

E a distância média entre dois grupos r e r' é dada por

$$\hat{C}_{rr'} = \frac{1}{n_r n_{r'}} \sum_{i=1}^{n_r} \sum_{j=1}^{n_{r'}} \phi(x_i^r, x_j^{r'}), \tag{3.7}$$

$$E(\hat{C}_{rr'}) = C(P_r, P_{r'}) = \int \int \phi(x, y) dP_r(x) dP_{r'}(y).$$

3.3.1 Decomposição em distância *Intra* e *Entre* grupos

A distância média geral ou soma de quadrados total (SQT) pode ser decomposta em função das quantidades (3.6) e (3.7), da seguinte maneira

$$\begin{aligned} SQT &= \binom{n}{2}^{-1} \left(\sum_{r=1}^{R} \binom{n_r}{2} \hat{Q}_r + \sum_{r<r'} n_r n_{r'} \hat{C}_{rr'} \right) \\ &= \sum_{r=1}^{R} \frac{n_r}{n} \hat{Q}_r + \sum_{r \neq r'} \frac{n_r n_{r'}}{n(n-1)} \hat{C}_{rr'} - \sum_{r=1}^{R} \frac{n_r (n - n_r)}{n(n-1)} \hat{Q}_r \\ &= SQI + SQE, \end{aligned} \tag{3.8}$$

em que SQI é a Soma de Quadrados Intra-grupos

$$SQI = \sum_{r=1}^{R} \frac{n_r}{n} \hat{Q}_r, \tag{3.9}$$

e SQE a Soma de Quadrados Entre-grupos, dada por

$$\begin{aligned} SQE &= \sum_{r \neq r'} \frac{n_r n_{r'}}{n(n-1)} \hat{C}_{rr'} - \sum_{r=1}^{R} \frac{n_r(n-n_r)}{n(n-1)} \hat{Q}_r \\ &= \frac{1}{n(n-1)} \left\{ \sum_{r=1}^{R-1} \sum_{r'=r+1}^{R} n_r n_{r'} (2\hat{C}_{rr'} - \hat{Q}_r - \hat{Q}_{r'}) \right\}. \end{aligned} \tag{3.10}$$

Então,

$$SQE = \frac{1}{n(n-1)} \left\{ \sum_{r=1}^{R-1} \sum_{r'=r+1}^{R} n_r n_{r'} (\hat{D}_{rr'}) \right\}, \tag{3.11}$$

e portanto,

$$E(SQE) = \frac{1}{n(n-1)} \left\{ \sum_{r=1}^{R-1} \sum_{r'=r+1}^{R} n_r n_{r'} (D_{rr'}) \right\}, \tag{3.12}$$

ou seja, a dissimilaridade média dos R grupos.

3.3.2 Teste de Hipóteses

Como o objetivo é avaliar se existe homogeneidade entre grupos com relação a uma determinada característica, a partir dos resultados apresentados nas equações (3.11) e (3.12) pode-se pensar no seguinte teste de hipóteses

$$\begin{aligned} H_0 : C(P_r, P_{r'}) &= \frac{1}{2}\left[Q(P_r) + Q(P_{r'})\right] \Rightarrow \\ 2C(P_r, P_{r'}) &= Q(P_r) + Q(P_{r'}), \\ H_a : 2C(P_r, P_{r'}) &> Q(P_r) + Q(P_{r'}), \end{aligned} \tag{3.13}$$

para todo $r = \{1, 2, ..., R\}$.

Portanto, sob H_0, $E(SQE) = 0$, o que implica que $E(SQT) = E(SQI)$. A estatística do teste será, portanto, a SQE descrita em (3.10). Resta agora encontrar a distribuição assintótica dessa estatística.

A estatística SQE é uma combinação linear de U-Estatísticas que possuem distribuição assintoticamente normais, obtidas a partir dos Teoremas 2.4 e 2.5, e portanto, ela também possui distribuição normal assintótica. Basta encontrar $Var(SQE)$.

$$\begin{aligned} Var(SQE) &= Var\left[\frac{1}{n(n-1)}\left(\sum_{r<r'} n_r n_{r'} 2\hat{C}_{rr'} - \sum_{r=1}^{R} n_r(n-n_r)\hat{Q}_r\right)\right] \\ &= Var\left(\sum_{r<r'} \frac{n_r n_{r'}}{n(n-1)} 2\hat{C}_{rr'}\right) + Var\left(\sum_{r=1}^{R} \frac{n_r(n-n_r)}{n(n-1)}\hat{Q}_r\right) + \\ &- 2Cov\left(\sum_{r<r'} \frac{n_r n_{r'}}{n(n-1)} 2\hat{C}_{rr'}, \sum_{r=1}^{R} \frac{n_r(n-n_r)}{n(n-1)}\hat{Q}_r\right). \end{aligned} \tag{3.14}$$

Calculando cada parcela separadamente, tem-se que

$$\begin{aligned} Var\left(\sum_{r<r'} \frac{n_r n_{r'}}{n(n-1)} 2\hat{C}_{rr'}\right) &= Cov\left(\sum_{r<r'} \frac{n_r n_{r'}}{n(n-1)} 2\hat{C}_{rr'}, \sum_{k<k'} \frac{n_k n_{k'}}{n(n-1)} 2\hat{C}_{kk'}\right) \\ &= \frac{4}{n^2(n-1)^2}\left[\sum_{r=1}^{R} n_r^2 Var(\hat{C}_{rr'})\right. \\ &+ \sum_{r<r'}\sum_{\substack{k'=r+1 \\ k'\neq r'}}^{R} Cov(\hat{C}_{rr'}, \hat{C}_{rk'}) + \\ &+ \left.\sum_{r<r'}\sum_{k\neq r}^{R-1} Cov(\hat{C}_{rr'}, \hat{C}_{kr'})\right]; \end{aligned} \tag{3.15}$$

$$Var\left(\sum_{r=1}^{R} \frac{n_r(n-n_r)}{n(n-1)}\hat{Q}_r\right) = \frac{1}{n^2(n-1)^2}\sum_{r=1}^{R} n_r^2(n-n_r)^2 Var(\hat{Q}_r), \tag{3.16}$$

$Cov(\hat{Q}_r, \hat{Q}_{r'}) = 0 \ \forall \ r \neq r'$ pois são provenientes de amostras distintas e independentes.

$$\begin{aligned} &Cov\left(\sum_{r<r'} \frac{n_r n_{r'}}{n(n-1)} 2\hat{C}_{rr'}, \sum_{k=1}^{R} \frac{n_k(n-n_k)}{n(n-1)}\hat{Q}_k\right) = \\ &= \frac{2}{n^2(n-1)^2}\left[\sum_{r<r'} n_r n_{r'} n_r(n-n_r) Cov(\hat{C}_{rr'}, \hat{Q}_r) + \right. \\ &+ \left.\sum_{r<r'} n_r n_{r'} n_{r'}(n-n_{r'}) Cov(\hat{C}_{rr'}, \hat{Q}_{r'})\right]. \end{aligned} \tag{3.17}$$

Defina-se as quantidades

1. $\phi_1(x_{r1}) = E[\phi(X_{r1}, X_{r2})|X_{r1} = x_{r1}]$;

2. $\psi_{0,1}^{rr'}(x_{r'1}) = E[\phi(X_{r1}, X_{r'1})|X_{r'1} = x_{r'1}]$;

3. $\psi_{1,0}^{rr'}(x_{r1}) = E[\phi(X_{r1}, X_{r'1})|X_{r1} = x_{r1}]$.

A partir dos Teoremas 2.2, 2.3 e 2.5 tem-se que

$$
\begin{aligned}
Var(\hat{Q}_r) &= \frac{4\sigma_1^2}{n_r}; && (3.18)\\
Var(\hat{C}_{rr'}) &= \frac{1}{n_{r'}}\sigma_{0,1}^2 + \frac{1}{n_r}\sigma_{1,0}^2; && (3.19)\\
Cov(\hat{C}_{rr'}, \hat{Q}_r) &= \frac{2}{n_r^2}Cov(\phi_1(X_{r1}), \psi_{1,0}^{rr'}(X_{r1})); && (3.20)\\
Cov(\hat{C}_{rr'}, \hat{Q}_{r'}) &= \frac{2}{n_{r'}^2}Cov(\phi_1(X_{r'1}), \psi_{0,1}^{rr'}(X_{r'1})); && (3.21)\\
Cov(\hat{C}_{rr'}, \hat{C}_{kr'}) &= \frac{1}{n_{r'}}Cov(\psi_{0,1}^{rr'}(X_{r'1}), \psi_{0,1}^{kr'}(X_{r'1})); && (3.22)\\
Cov(\hat{C}_{rr'}, \hat{C}_{rk'}) &= \frac{1}{n_r}Cov(\psi_{1,0}^{rr'}(X_{r1}), \psi_{1,0}^{rk'}(X_{r1})). && (3.23)
\end{aligned}
$$

em que $\sigma_1^2 = Var(\phi_{(1)}(X_{r1}))$, $\sigma_{0,1}^2 = Var(\psi_{0,1}^{rr'}(X_{r'1}))$ e $\sigma_{1,0}^2 = Var(\psi_{1,0}^{rr'}(X_{r1}))$.

E portanto, no caso em que $R = 2$,$Var(SQE)$ será dada por

$$
Var(SQE) = \frac{n_1^2 n_2^2}{n^2(n-1)^2}\left[4Var(\hat{C}_{12}) + Var(\hat{Q}_1) + Var(\hat{Q}_2)\right] + O\left(\frac{1}{n}\right).
$$

Pinheiro, Pinheiro e Sen (2008) mostraram que, sob certas condições, a Soma de Quadrados Entre grupos sob a hipótese nula (de homogeneidade entre os grupos) tem a seguinte distribuição

$$
\frac{nSQE}{\sqrt{\binom{n}{2}U_n^{(4)}}} \to N(0,1),
$$

$$
nSQE \to N(0, 2\xi_0)
$$

e

$$
\binom{n}{2}U_n^{(4)} \to \xi_0,
$$

em que, $U_n^{(4)} = \sum_{1\le i<j\le n}\phi_{(2)}^2(X_i, X_j)$.

A partir desse resultado pode se pensar em calcular o poder do teste de hipótese, ou seja, a probabilidade de rejeitar a hipótese nula quando ela é falsa. Note que, sob H_1, o valor esperado de SQE, é dado por

$$\begin{aligned} E_{H_1}[SQE] &= \frac{1}{n(n-1)} \sum_{g<g'} n_g n_{g'} (2\hat{C}_{gg'} - \hat{Q}_g - \hat{Q}_{g'}) \\ &= \sum_{g<g'} \frac{n_g}{n} \frac{n_{g'}}{n-1} (2\hat{C}_{gg'} - \hat{Q}_g - \hat{Q}_{g'}) \\ &\overset{n\to\infty}{\longrightarrow} \sum_{g<g'} p_g p_{g'} (2\hat{C}_{gg'} - \hat{Q}_g - \hat{Q}_{g'}) \equiv \theta_1, \end{aligned}$$

para todo $g, g' = 1, 2, ..., G$. Então, $E_{H_1}[SQE] = \theta_1 + O(n^{-2})$ e $E_{H_1}[nSQE] = n\theta_1 + O(n^{-1})$.

Seja $\theta_1 = \delta_n \equiv \Delta/n$. Então, $\delta_n \to 0$ quando $n \to \infty$, e $E[nSQE] = O(1)$. Observe então a hipótese $2C_{gg'} - Q_g - Q_{g'} = o(n^{-1})$, $1 \leq g \neq g' \leq G$.

$$\begin{aligned} P(H_0 \; ser \; rejeitada | H_1) &= P\left(\frac{nSQE}{\sqrt{\binom{n}{2} U_n^{(4)}}} \geq q_\alpha | H_1 \right) \\ &= P\left(\frac{nSQE - n\theta_1}{\sqrt{\binom{n}{2} U_n^{(4)}}} \geq q_\alpha - \frac{n\theta_1}{\sqrt{\binom{n}{2} U_n^{(4)}}} \right) \end{aligned}$$

onde $\sigma^2 = lim_{n\to\infty} \binom{n}{2} [E_{H_1}(SQE^2 - \theta_1^2)]$ e

$$P_{H_0}\left(\frac{nSQE}{\sqrt{\binom{n}{2} U_n^{(4)}}} \geq q_\alpha \right) \to \alpha$$

quando $n \to \infty$.

Mas,

$$\frac{nSQE}{\sqrt{\binom{n}{2} U_n^{(4)}}} \xrightarrow{\mathcal{D}} N(\theta_1, \sigma^2)$$

Finalmente,

$$\begin{aligned} P_{H_1}\left(\frac{nSQE}{\sqrt{\binom{n}{2} U_n^{(4)}}} > q_\alpha \right) &= P_{H_1}\left(\frac{n(SQE - \theta_1)}{\sqrt{\binom{n}{2} U_n^{(4)}}} > q_\alpha - \frac{n\theta_1}{\sqrt{\binom{n}{2} U_n^{(4)}}} \right) \\ &\longrightarrow 1 - \Phi\left(q_\alpha - \frac{\Delta}{\sigma} \right). \end{aligned} \tag{3.24}$$

Nesta seção foi investigado o comportamento de SQE para testar a hipótese definida em (3.13). Sob a hipotese nula de homogeneidade ou pelas alternativas de Pitman, SQE tem uma distribuição assintoticamente normal. O poder do teste para as alternativas de Pitman pode ser derivado de (3.24).

3.4 O problema de multifatores

Até agora foram apresentados resultados para dados com uma classificação, correspondente a análise de variância clássica ("*one-way ANOVA*", em que as populações são identificadas pelos níveis de um único fator). Nayak e Gastwirth (1989) abordam o problema de multifatores da seguinte maneira.

Considere dois fatores, A_1 com s níveis e A_2 com t níveis (a teoria é facilmente expandida para os casos de mais de 2 fatores). Seja P_{ij}, a função de probabilidade da variável aleatória X com relação ao $i-ésimo$ nível de A_1 e ao $j-ésimo$ nível de A_2, e seja λ_{ij} a fração da população de interesse pertencente a este subgrupo. Para obter o efeito conjunto de A_1 e A_2, considera-se a classificação cruzada de A_1 e A_2, como um fator único com $s \times t$ níveis, obtendo a partir disso a decomposição da SQT, da seguinte forma

$$SQT = SQI(A_1, A_2) + SQE(A_1, A_2), \tag{3.25}$$

onde $SQT = Q(\sum\sum \lambda_{ij} P_{ij})$, $SQI(A_1, A_2) = \sum\sum \lambda_{ij} Q(P_{ij})$ e $SQE(A_1, A_2) = SQT - SQI(A_1, A_2)$.

Como na partição da soma de quadrados na análise de regressão, $SQE(A_1, A_2)$ pode ser decomposta da seguinte forma

$$SQE(A_1, A_2) = SQE(A_1) + SQE(A_2|A_1), \tag{3.26}$$

em que, $SQE(A_1) = SQT - \sum \lambda_{i\cdot} Q(P_{i\cdot})$, $\lambda_{i\cdot} = \sum_j \lambda_{ij}$, $P_{i\cdot} = \sum_j (\lambda_{ij}/\lambda_{i\cdot}) P_{ij}$ e $SQE(A_2|A_1) = SQE(A_1, A_2) - SQE(A_1)$. Para um dado nível de A_1, dito $A_1 = i$, as desigualdades entre os níveis de A_2 são

$$SQE(A_2|A_1 = i) = Q(P_{i\cdot}) - \sum_j \frac{\lambda_{ij}}{\lambda_{i\cdot}} Q(P_{ij}). \tag{3.27}$$

E pode-se notar então que

$$SQE(A_2|A_1) = \sum_i \lambda_i . SQE(A_2|A_1 = i). \tag{3.28}$$

Portanto $SQE(A_2|A_1)$ é uma média ponderada das desigualdades entre os níveis de A_2 para cada nível de A_1. Isso representa a proporção da variabilidade não explicada por A_1 que é explicada por A_2.

A metodologia pode ser generalizada para cobrir múltiplos fatores. SQT é sempre definida como a variabilidade na população total, que é uma mistura de muitas sub-populações. Quando há k fatores, $A_1, ..., A_k$, $SQI(A_1, ..., A_k)$ é a média ponderada das desigualdades dentro de cada grupo definido pela classificação cruzada de $A_1, ..., A_k$ e $SQE(A_1, ..., A_k) = SQT - SQI(A_1, ..., A_k)$. Analogamente,

$$SQE(A_1, ..., A_s|A_{s+1}, ..., A_k) = SQE(A_1, ..., A_k) - SQE(A_{r+1}, ..., A_k). \tag{3.29}$$

Portanto,

- $SQE(A_i)$ pode ser interpretado como o efeito principal do fator A_i para todo $k = 1, 2, ..., K$;
- $SQE(A_1, ..., A_s)$ é o efeito de interação, ou efeito conjunto, entre os fatores $A_1, ..., A_s$ $(s \neq k)$;
- $SQE(A_i|A_j)$ é o efeito do fator A_i condicionado ao fator A_j, ou seja, é o efeito do fator A_i retirando-se o efeito do fator A_j.

Capítulo 4

Metodologias

Neste capítulo será apresentada uma descrição mais detalhada das duas metodologias que são propostas.

4.1 Metodologia 1 - Diversidade baseada no ganho relativo

A primeira metodologia proposta se baseia no uso da variável *ganho relativo*. O objetivo é avaliar, dado $g = 1, 2, ..., G$ grupos, se há diferença entre eles com relação a essa medida de desempenho. Os principais grupos de interesse nesse estudo são os formados pelo tipo de escola cursada no segundo grau (particular ou pública) e os formados pelo sexo.

Com base nas medidas de diversidade propostas por Rao (1982), apresentadas no Capítulo 4, foi tomada como função núcleo a diferença quadrática, ou seja, $\phi(x, y) = (x - y)^2$.

A medida de diversidade entre os grupos (DIV) será dada por

$$Q_g = \int\int (x_{g1} - x_{g2})^2 dP_g(x_{g1}) dP_g(x_{g2}),$$

em que X_{gi} é o ganho relativo do $i - ésimo$ aluno do grupo g.

A medida de dissimilaridade (DIV) entre os grupos é então,

$$C_{gg'} = \int\int (X_{g1} - X_{g'1})^2 dP_g(x_{g1}) dP_{g'}(x_{g'1}).$$

Os estimadores dessas quantidades são baseados em U-Estatísticas

$$\hat{Q}_g = \binom{n_g}{2}^{-1} \sum_{i<j} (x_{gi} - x_{gj})^2$$

e

$$\hat{C}_{gg'} = \frac{1}{n_g n_{g'}} \sum_i \sum_j (x_{gi} - x_{g'j})^2.$$

A Soma de Quadrados Entre grupos, SQE, que será utilizada como estatística do teste é então

$$SQE = \binom{n}{2}^{-1} \sum_{g<g'} n_g n_{g'} \left(2\hat{C}_{gg'} - \hat{Q}_g - \hat{Q}_{g'}\right).$$

Como a função ϕ é um medida euclidiana, então ela atende as quesitos do Teorema 3.1 e é verdadeiro que $C_{gg'} \geq \frac{1}{2}(Q_g + Q_{g'})$ (valendo a igualdade quando há homogeneidade entre os grupos testados). Então pode se construir o seguinte teste de hipóteses

$$\begin{aligned} H_0 &: 2C_{gg'} - Q_q - Q_{g'} = 0 \\ H_a &: 2C_{gg'} - Q_q - Q_{g'} > 0 \end{aligned}$$

para todo $g = 1, ..., G$.

4.1.1 Calculando as variâncias dos estimadores

Sejam x_{11} , x_{12} , ... , x_{1n_1} , x_{21} , x_{22} , ... , x_{2n_2} , x_{G1} , x_{G2} , ... , x_{Gn_G}, G amostras aleatórias, obtidas de distribuições $F(\mu_g, \sigma_g)$, respectivamente, para $g = 1, ..., G$. Seja $n = \sum_{g=1}^{G} n_g$.

A estatística $\hat{Q}_g$ é uma U-Estatística de grau 2, onde $\phi(x_{g1}, x_{g2}) = (x_{g1} - x_{g2})^2$ para todo $g = 1, ..., G$. Então,

$$\phi_1(x_{g1}) = E[\phi(X_{g1}, X_{g2}) | X_{g1} = x_{g1}]$$

$$\begin{aligned} &= E[(X_{g1} - X_{g2})|X_{g1} = x_{g1}] \\ &= E[x_{g1}^2 - 2x_{g1}X_{g2} + X_{g2}^2] \\ &= x_{g1}^2 - 2x_{g1}E[X_{g2}] + E[X_{g2}^2] \\ &= x_{g1}^2 - 2x_{g1}\mu_g + \mu_g^2, \end{aligned} \tag{4.1}$$

e portanto,

$$\begin{aligned} E[\phi_1(X_{g1})] &= E[X_{g1}^2 - 2X_{g1}\mu_g + \mu_g^2] \\ &= \mu_g^2 - 2\mu_g\mu_g + \mu_g^2 \\ &= 2\mu_g^2 - 2(\mu_g)^2 \\ &= 2\sigma_g^2 \\ &= \theta(F_g). \end{aligned} \tag{4.2}$$

Para o cálculo da $Var(\phi_1)$ é preciso ainda calcular $E[\phi_1^2]$, que é dado por

$$\begin{aligned} E[\phi_1^2] &= E[X_{g1}^2 - 2X_{g1}\mu_g + \mu_g^2]^2 \\ &= E[X_{g1}^4 - 2X_{g1}^3\mu_g + X_{g1}^2\mu_g^2 - 2X_{g1}^3\mu_g + 4X_{g1}^2(\mu_g)^2 \\ &- 2X_{g1}\mu_g\mu_g^2 + X_{g1}^2\mu_g^2 - 2X_{g1}\mu_g\mu_g^2 + (\mu_g^2)^2] \\ &= E[X_{g1}^4 - 4X_{g1}^3\mu_g + 2X_{g1}^2\mu_g^2 + 4X_{g1}^2(\mu_g)^2 - 4X_{g1}\mu_g\mu_g^2 + (\mu_r^2)^2] \\ &= \mu_g^4 - 4\mu_g^3\mu_g + 2(\mu_g^2)^2 + 4\mu_g^2(\mu_g)^2 - 4\mu_g^2(\mu_g)^2 + (\mu_g^2)^2 \\ &= \mu_g^4 - 4\mu_g^3\mu_g + 3(\mu_g^2)^2, \end{aligned} \tag{4.3}$$

e daí segue que,

$$Var(\phi(X_{g1})) = \mu_g^4 - 4\mu_g^3\mu_g + 3(\mu_g^2)^2 - 4(\sigma_g^2)^2. \tag{4.4}$$

E portanto, do Teorema 2.4 da convergência assintótica para U-Estatísticas tem-se que

$$\sqrt{n_g}(\hat{Q}_g - \theta(F_g)) \xrightarrow{D} N\left(0, 4Var(\phi(X_{g1}))\right). \tag{4.5}$$

A estatística $\hat{C}_{gg'}$ é uma U-Estatística bi-dimensional de grau $(1,1)$, para todo $g, g' = 1, ..., G$.

$$C_{gg'} = \int\int (x_g - x_{g'})^2 dF_g(x_g) dF_{g'}(x_{g'})$$

$$
\begin{aligned}
&= E[X_g - X_{g'}]^2 \\
&= E[(X_g)^2 - 2X_gX_{g'} + X_{g'}^2] \\
&= E\left\{E[X_g^2 - 2X_gX_{g'} + X_{g'}^2]|X_g = x_g\right\} \\
&= E\left\{x_g^2 - 2x_gE(X_{g'}) + E(X_{g'}^2)\right\} \\
&= E[X_g^2 - 2X_g\mu_{g'} + \mu_{g'}^2] \\
&= \mu_g^2 - 2\mu_g\mu_{g'} + \mu_{g'}^2 \\
&= \theta(F_g, F_{g'}). \qquad (4.6)
\end{aligned}
$$

Calcula-se então o seguinte,

$$
\begin{aligned}
\Phi_{10}(x_g) &= E[\phi(X_g, X_{g'})|X_g = x_g] \\
&= E[x_g^2 - 2x_gX_{g'} + X_{g'}^2] \\
&= x_g^2 - 2x_g\mu_{g'} + \mu_{g'}^2. \qquad (4.7)
\end{aligned}
$$

Da mesma forma,

$$
\begin{aligned}
\Phi_{01}(x_{g'}) &= E[\phi(X_g, X_{g'})|X_{g'} = x_{g'}] \\
&= \mu_g^2 - x_{g'}\mu_g + x_{g'}^2. \qquad (4.8)
\end{aligned}
$$

Logo,

$$
\begin{aligned}
\varsigma_{10} &= E[\Phi_{10}(X_g) - C_{gg'}^2] \\
&= E\left[\left(X_g^2 - 2X_g\mu_{g'} + \mu_{g'}^2\right)^2 - C_{gg'}^2\right] \\
&= E[X_g^4 - 2X_g^3\mu_{g'} + X_g^2\mu_{g'}^2 - 2X_g^3\mu_{g'} + 4X_g^2\mu_{g'}^2 - 2X_g\mu_{g'}\mu_{g'}^2 + X_g^2\mu_{g'}^2 \\
&- 2X_g\mu_{g'}\mu_{g'}^2 + (\mu_{g'})^2] - C_{gg'}^2 \\
&= E\left[X_g^4 - 4X_g^3\mu_{g'} + 2X_g^2\mu_{g'}^2 + 4X_g^2(\mu_{g'})^2 - 4X_g\mu_{g'}\mu_{g'}^2 + (\mu_{g'})^2\right] - C_{gg'}^2 \\
&= \mu_g^4 - 4\mu_g^3\mu_{g'} + 2\mu_g^2\mu_{g'}^2 - 4\mu_g\mu_{g'}\mu_{g'}^2 + (\mu_{g'}^2)^2 - C_{gg'}^2. \qquad (4.9)
\end{aligned}
$$

Analogamente,

$$
\begin{aligned}
\varsigma_{01} &= E\left[\Phi_{01}(X_{r'}) - C_{gg'}^2\right] \\
&= \mu_{g'}^4 - 4\mu_{g'}^3\mu_g + 2\mu_{g'}^2\mu_g^2 - 4\mu_{g'}\mu_g\mu_g^2 + (\mu_g^2)^2 - C_{gg'}^2. \qquad (4.10)
\end{aligned}
$$

E portanto,

$$Var(\hat{C}_{gg'}) = \frac{1}{n_g}\varsigma_{10} + \frac{1}{n_{g'}}\varsigma_{01}. \tag{4.11}$$

A variância da estatística do teste (SQE) é dada pela equação (3.14) e é função das variâncias das estatísticas $\hat{Q}_g$ e $\hat{C}_{gg'}$ e de suas co-variâncias.

4.1.2 Estimação das Variâncias

Para esta metodologia serão feitas duas abordagens, uma dita paramétrica e outra não paramétrica.

Como pode ser visto nas equações (5.4), (5.9), (5.10) e (5.11), as variâncias das estatísticas $\hat{Q}_g$ e $\hat{C}_{gg'}$ são funções dos momentos de ordem 1,2,3, e 4 (μ_1, μ_2, μ_3 *e* μ_4) da função de distribuição assumida para a variável em estudo, no caso o ganho relativo. na abordagem paramétrica, onde serão feitos dois estudos (um assumindo a distribuição normal e um segundo assumindo a distribuição triangular), serão estimados por máxima verossimilhança os momentos da distribuição e a partir dessas estimativas será calculada a variância estimada para a estatística do teste. Com a estimativa da variância irá se construir intervalos de confiança para a SQE.

Na segunda abordagem, dita não paramétrica, as variâncias e os intervalos de confiança serão calculados pelo método de re-amostragem de *jackknife*. Ela recebe este nome por não se assumir nenhuma função de distribuição para os dados.

Em ambas as abordagens o objetivo de se obter uma estimativa para variância de SQE é que a partir dela pode-se construir intervalos de $100(1-\alpha)\%$ de confiança, da seguinte forma

$$SQE \pm z_{\frac{\alpha}{2}}\hat{Var}(SQE). \tag{4.12}$$

em que $z_{\frac{\alpha}{2}}$ representa o valor tabelado da normal-padrão com área da curva abaixo igual a $1-\frac{\alpha}{2}$.

Uma vez construído um intervalo de $100(1-\alpha)\%$ de confiança, para avaliar se a estatística SQE é significativamente diferente de zero, ao nivel de $100\alpha\%$, basta olhar para

o intervalo de confiança. Se o intervalo não compreender o valor *zero* então a estatística é significativa, caso contrário, não será significativa (não rejeita-se a hipótese nula).

Essa ligação entre intervalos de confiança e testes de hipóteses, permitindo que, na prática, calcule-se o primeiro e tire-se conclusões sobre o segundo, é mostrado pelo seguinte resultado.

Proposição 4.1 (Dualidade entre Intervalos de Confiança e testes de Hipóteses)**.**

Sejam $x_1, ..., x_n$ observações de $X_1, ..., X_n$ i.i.d. F, $\theta \in \Theta$ um parâmetro real, T uma estatística e $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Considere um intervalo de $100(1-\alpha)\%$ de confiança, $(t_\alpha(\boldsymbol{X}), T_\alpha(\boldsymbol{X}))$, e uma região crítica com nível α, R_α e a região favorável à hipótese A_α. Então, pode-se construir $(t_\alpha(\boldsymbol{X}), T_\alpha(\boldsymbol{X}))$ a partir de A_α e vice-versa.

Utilizando a distribuição normal

Nesta abordagem assume-se que o ganho relativo, para cada um dos G grupos de interesse, venha de uma distribuição Normal de parâmetros μ_g e σ_g. Os estimadores de máxima verossimilhança para esses parâmetros são a média amostral ($\overline{X}$) e a variância amostral (S^2), respectivamente. Portanto, tem se que

$$\hat{\mu} = \overline{X} \; e \; \hat{\sigma}^2 = S^2.$$

A partir disso estima-se o segundo momento da seguinte forma

$$\hat{\mu_2} = S^2 + \overline{X}^2.$$

Para estimar os momentos de ordem 3 e 4 (μ_3 e μ_4), pode-se o seguinte resultado.

Lema 4.1 (Stein's Lemma)**.**

Seja X $N(\theta, \sigma^2)$, e seja g um função diferenciável que satisfaça $E|g'(X)| < \infty$. Então

$$E[g(X)(X-\theta)] = \sigma^2 E g'(X).$$

Portanto, pelo Lema 4.1, se X tem distribuição $N(\mu, \sigma^2)$, tem-se que

$$\begin{aligned}
\mu_3 &= EX^3 \\
&= EX^2(X - \mu + \mu) \\
&= EX^2(X - \mu) + \mu EX^2 \\
&= 2\sigma^2 EX + \mu EX^2 \\
&= 2\sigma^2\mu + \mu(\sigma^2 + \mu^2) \\
&= 3\mu\sigma^2 + \mu^3. \qquad (4.13)
\end{aligned}$$

De forma análoga, para se calcular μ_4 usa-se do mesmo artifício,

$$\begin{aligned}
\mu_4 &= EX^4 \\
&= EX^3(X - \mu + \mu) \\
&= EX^3(X - \mu) + \mu EX^3 \\
&= 3\sigma^2 EX^2 + \mu EX^3 \\
&= 3\sigma^2(\sigma^2 + \mu^2) + \mu(3\mu\sigma^2 + \mu^3) \\
&= 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4. \qquad (4.14)
\end{aligned}$$

Portanto, os estimadores para μ_3 e μ_4 são

$$\begin{aligned}
\hat{\mu}_3 &= 3\overline{X}S^2 + \overline{X}^3, \\
\hat{\mu}_4 &= 3S^4 + 6S^2\overline{X}^2 + \overline{X}^4.
\end{aligned} \qquad (4.15)$$

A partir das estimativas das variâncias de $\hat{Q}_g$ e $\hat{C}_{gg'}$, estima-se a variância da SQE e se calcula um intervalo de 95% confiança. As conclusões sobre rejeitar ou não a hipótese de homogeneidade entre os grupos é obtida analisando o intervalo de confiança como descrito anteriormente.

Utilizando a distribuição triangular

Uma segunda análise será feita utilizando a distribuição triangular. Para estimar os parâmetros da distribuição triangular pelo método de máxima verossimilhança será utilizado o software *MLE-Estimator*. Uma abordagem da distribuição triangular é apresentada no Apendice A.

Suponha que o ganho relativo, para cada grupo G, tenha distribuição triangular em $[a_g, b_g]$ e moda m_g. Obtém as estimativas $\hat{a}_g$, $\hat{b}_g$ e $\hat{m}_g$ pelo método de máximo verossimilhança. As estimativas dos momentos da distribuição podem ser obtidas a partir da equação (1.3).

Com as estimativas dos momentos μ_1,μ_2, μ_3 e μ_4, estima-se a variância de SQE como descrito na seção anterior e, calcula-se um intervalo de 95% de confiança.

Abordagem não paramétrica

Com o Teorema Central do Limite (TCL), os estimadores paramétricos (e não paramétricas como as U-Estatística) lineares têm, sob certas condições de regularidade, distribuição normal. Nesse caso, falta descobrir qual o valor da variância de um tal estimador. Em geral, não será garantida a obtenção analítica ou com boa aproximação do valor de σ. Uma opção computacionalmente intensiva é por técnicas de re-amostragem. Na análise não paramétrica do ganho relativo, e também na análise de diversidade das disciplinas, será utilizado o método de re-amostragem por *jackkinfe* que é um caso particular do *bootstrap*(Davison e Hinkley, 1999).

Definição 4.1 (Algoritmo *jackknife* para estimação do erro-padrão)**.**

Sejam $\boldsymbol{x}$ uma amostra de F e $s(\boldsymbol{x})$ uma estimativa de θ.

1. *Selecione n amostras jackknife $\boldsymbol{x}_{(-1)}, \boldsymbol{x}_{(-2)}, ..., \boldsymbol{x}_{(-n)}$, sendo $\boldsymbol{x}_{(-i)}$ um vetor (n-1)-dimensional, pela exclusão da $i-$ésima observação, $i = 1, ..., n$.*

2. *Calcule a replicação jackknife*

$$\hat{\theta}^*(-i) = s(\boldsymbol{x}_{(-i)}) \; i = 1, 2, ...n.$$

3. Calcule

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \frac{\hat{\theta}^*(-i)}{n}.$$

4. Estime o erro-padrão de $s(\boldsymbol{x})$ por

$$\hat{ep}_{JACK} = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)})^2 \right].$$

A partir deste algoritmo estima-se a variância da SQE ($\hat{Var}(SQE)_{JACK}$) e calcula-se o intervalo de 95% de confiança como descrito em (4.12).

4.2 Metodologia 2 - Utilizando as notas das disciplinas

Como dito anteriormente, a nota final do vestibular reflete apenas o desempenho do aluno no ingresso à universidade em anos diferentes, não sendo possível avaliar seu desenvolvimento durante o curso apenas com esta variável. O coeficiente de rendimento final (CR), não é comparável entre alunos de turmas diferentes, pois cada curso dispõe de metodologias distintas de avaliação e, portanto, não é coerente comparar o CR de um aluno da medicina com um aluno da matemática, por exemplo. Essas questões foram amplamente exploradas no Capítulo 1.

Com o objetivo de propor métodos mais robustos e com a dificuldade em se usar a nota do vestibular e o CR, pensou-se em utilizar as notas das disciplinas cursadas pelos estudantes. De forma que, para se avaliar, entre dois alunos $\mathbb{A}$ e $\mathbb{B}$, quem teve melhor desempenho, olha-se para o grupo de disciplinas cursadas em comum entre ambos e observa-se, por exemplo, a proporção das disciplinas em que $\mathbb{A}$ obteve melhores resultados do que $\mathbb{B}$. Isso condicionado ao fator "nota do vestibular", ou seja, condicionado ao posto do aluno com relação a nota final do vestibular.

É importante lembrar que, uma vez que uma determinada disciplina pode ser ministrada por professores difentes em períodos distintos e até mesmo num mesmo período, os métodos de avalição não são os mesmos. Entretanto, para esta metodologia, irá se supor

que não há diferença entre os métodos de avaliação quando se compara dois alunos com relação ao desempenho em uma mesma discplina.

Como na prática um aluno pode cursar uma mesma disciplina por várias vezes, no caso dele ser reprovado, para determinar se um aluno foi melhor do que o outro numa determinada disciplina, duas características foram avaliadas: a nota obtida na disciplina e o número de vezes que o aluno fez a mesma. No caso em que ambos os alunos fizeram determinada disciplina o mesmo número de vezes é considerado "o melhor" aquele que obteve maior média, caso contrário, é considerado "o melhor" aquele que fez o menor número de vezes a disciplina.

Sejam $\mathbf{Y}_{a1}, ..., \mathbf{Y}_{an_a}$ vetores com as notas das disciplinas cursadas pelos alunos ingressantes no ano a, ou seja, $\mathbf{Y}_{ai} = (Y_{ai1}, ..., Y_{aik_i})$, em que k_i é o número de disciplinas cursadas pelo $i-ésimo$ aluno ingressante no ano a e i representa a posição do aluno no ano a com relação a nota obtida no vestibular ($i = 1$ representa o aluno com melhor nota, e $i = n_a$ o aluno com pior nota).

Assume-se que Y_{ail} tem um determinada distribuição $F_{al}(\mu_{al}, \sigma^2_{al})$, para $l = 1, 2, ..., k_i$, distribuição esta desconhecida.

A função $\phi(\cdot)$ é definida da seguinte forma

$$\phi(Y_{ail}, Y_{ajl}|i < j) = I(Y_{ail} < Y_{ajl}), \tag{4.16}$$

onde l denota a disciplina feita em comum entre o $i-ésimo$ e o $j-ésimo$ aluno.

$$\begin{aligned} E\left[\phi(Y_{til}, Y_{tjl}|i < j)\right] &= P(Y_{ail} < Y_{ajl}|i < j) \\ &= \int_0^{10} \int_{y_{ail}}^{10} dF(y_{ajl}|i > j) dF(y_{ail}|i > j) \end{aligned} \tag{4.17}$$

(considerando que as notas variam entre 0 e 10). Ou seja, o valor esperado de $\phi(Y_{til}, Y_{tjl}|i < j)$, $E\left[\phi(Y_{til}, Y_{tjl}|i < j)\right]$, ***é a probabilidade do*** $j-ésimo$ ***ingressante no ano*** a ***ter tirado nota maior na disciplina*** l ***comparado ao*** $i-ésimo$ ***aluno, condicionado ao fato de que o*** $i-ésimo$ ***aluno obteve melhor desempenho no vestibular***.

A Soma de Quadrados Total no ano a será dada por

$$\begin{aligned} SQT_a &= \binom{n_a}{2}^{-1} \sum_{i \neq j} \sum_{l=1}^{k_{ij}} \frac{1}{k_{ij}} \phi\left(Y_{ail}, Y_{ajl} | i < j\right) \\ &= \binom{n_a}{2}^{-1} \sum_{i \neq j} \sum_{l=1}^{k_{ij}} \frac{1}{k_{ij}} I(Y_{ail} < Y_{ajl}), \end{aligned} \tag{4.18}$$

onde k_{ij} é o número de disciplinas cursadas em comum entre o $i-ésimo$ e o $j-ésimo$ aluno.

E a Soma de Quadrados Total para todos os anos $a = 1, 2, ..., A$ será

$$\hat{SQT} = \frac{1}{A} \sum_{a=1}^{A} \hat{Q}_{ga}. \tag{4.19}$$

No caso de G grupos e $\sum_{g=1}^{G} n_{ag} = n_a$, o estimador da variabilidade dentro do grupo g é definido como

$$\hat{Q}_{ag} = \binom{n_{ag}}{2}^{-1} \sum_{i \neq j} \sum_{l=1}^{k_{ij}} \frac{1}{k_{ij}} \phi\left(Y^g_{ail}, Y^g_{ajl} | i < j\right). \tag{4.20}$$

De maneira que a variabilidade total dentro de cada grupo seja obtida pela média da variabilidade dentro de cada grupos n ano $a = 1, 2, .., A$.

$$\hat{Q}_g = \frac{1}{A} \sum_{a=1}^{A} \hat{Q}_{ag}. \tag{4.21}$$

O valor esperado de $\hat{Q}_g$ é dado por Q_g para $g = 1, 2, ..., G$. Esta quantidade pode ser interpretada como a ***probabilidade de um aluno com pior desempenho no vestibular obter melhores notas nas disciplinas cursadas na universidade comparados a um aluno com melhor desempenho no vestibular, sendo ambos pertencentes a um mesmo grupo*** g.

Antes de se obter o estimador da variabilidade entre os grupos g e g' ($\hat{C}_{gg'}$) defina-se duas quantidades,

$$\hat{C}^*_{agg'} = \frac{1}{n_{ag} n_{ag'}} \left(\sum_{i \neq j} \sum_{l=1}^{k_{ij}} \phi(Y^g_{Ail}, Y^{g'}_{Ajl} | i < j) \right), \tag{4.22}$$

e

$$\hat{C}^*_{ag'g} = \frac{1}{n_{ag}n_{ag'}} \left(\sum_{i \neq j} \sum_{l=1}^{k_{ij}} \phi(Y^{g'}_{ajl}, Y^{g}_{ail} | j < i) \right); \tag{4.23}$$

ou seja, $C^*_{agg'}$ é a ***probabilidade de um aluno pertencente ao grupo*** g'***, que obteve desempenho inferior ao de um aluno pertencente ao grupo*** g ***no vestibular, ter obtido notas melhores durante a graduação, ambos ingressantes no ano*** a. Analogamente, $C^*_{ag'g}$ é a ***probabilidade de um aluno pertencente ao grupo*** g***, que obteve desempenho inferior ao de um aluno pertencente ao grupo*** g' ***no vestibular, ter obtido notas melhores durante a graduação***.

Se $n^*_{agg'}$ denotar o número de comparações feitas em $C^*_{agg'}$ e $n^*_{ag'g}$ o número de comparações feitas em $C^*_{ag'g}$. Então $n^*_{agg'} + n^*_{ag'g} = n_{ag}n_{ag'}$, isto é, o número total de comparações feitas entre os grupos g e g'. Desta forma

$$\hat{C}_{agg'} = \frac{n^*_{agg'}}{n_{ag}n_{ag'}} C^*_{agg'} + \frac{n^*_{ag'g}}{n_{ag}n_{ag'}} C^*_{ag'g}, \tag{4.24}$$

é a variabilidade total entre dois grupos para o ano a.

Assim, $\hat{C}_{gg'}$ é obtido a partir da seguinte equação

$$\hat{C}_{gg'} = \frac{1}{A} \sum_{a=1}^{A} \hat{C}_{agg'}. \tag{4.25}$$

A quantidade $C_{gg'}$, que é a $E(\hat{C}_{gg'})$, para $g, g' = 1, 2, ..., G$, pode ser interpretada como a ***probabilidade de um aluno pertencente ao grupo g (g') que obteve um pior desempenho no vestibular, ter notas melhores nas disciplinas cursadas durante a graduação, comparado a um aluno pertencente ao grupo g' (g) que obteve melhor desempenho no vestibular***.

A partir das equações (4.20),(4.24), pode-se obter a seguinte decomposição para a variabilidade Total(SQT).

$$\begin{aligned} SQT_a &= \binom{n_a}{2}^{-1} \left(\sum_{g=1}^{G} \binom{n_{ag}}{2} \hat{Q}_{ag} + \sum_{g<g'} n_{ag}n_{ag'} \hat{C}_{agg'} \right) \\ &= \sum_{g=1}^{G} \frac{n_{ag}}{n_a} \hat{Q}_{ag} + \sum_{g \neq g'} \frac{n_{ag}n_{ag'}}{n_a(n_a-1)} \hat{C}_{agg'} - \sum_{g=1}^{G} \frac{n_{ag}(n_a - n_{ag})}{n_a(n_a-1)} \hat{Q}_{ag} \end{aligned} \tag{4.26}$$

sendo,

$$SQI_a = \sum_{g=1}^{G} \frac{n_{ag}}{n_a} \hat{Q}_{ga} \tag{4.27}$$

e

$$\begin{aligned} SQE_a &= \sum_{g \neq g'} \frac{n_{ag} n_{ag'}}{n_a(n_a-1)} \hat{C}_{gg'a} - \sum_{g=1}^{G} \frac{n_{ag}(n-n_{ag})}{n_a(n_a-1)} \hat{Q}_{ga} \\ &= \frac{1}{n(n-1)} \sum_{g<g'} n_{ag} n_{ag'} (2\hat{C}_{agg} - \hat{Q}_{ag} - \hat{Q}_{ag}). \end{aligned} \tag{4.28}$$

A SQE geral é dada pela média aritmética dos SQE_a.

Como a função $\phi(\cdot)$ é assimétrica, o resultado $2C_{agg'} \geq Q_{ag} + Q_{ag'}$ não é verdadeiro, pois não satisfaz as condições do Teorema 3.1. Portanto, um teste de hipótese adequado seria

$$\begin{aligned} H_0 &: 2C_{gg'} - Q_g - Q_{g'} = 0 \\ H_a &: 2C_{gg'} - Q_g - Q_{g'} \neq 0. \end{aligned}$$

para todo $g = 1, ..., G$.

A partir da equação (4.24), para o caso de apenas 2 grupos, a soma de quadrados entre grupos pode ser decomposta como se segue

$$\begin{aligned} SQE &= \frac{1}{4} \sum_{a=1}^{A} \frac{1}{n_a(n_a-1)} n_{a1} n_{a2} (2\hat{C}_{a12} - \hat{Q}_{a1} - \hat{Q}_{a2}) \\ &= \frac{1}{4} \sum_{a=1}^{A} \frac{1}{n_a(n_a-1)} n_{a1} n_{a2} \left(2 \left(\frac{n^*_{a12}}{n_{a1} n_{a2}} \hat{C}^*_{a12} + \frac{n^*_{a21}}{n_{a1} n_{a2}} \hat{C}^*_{a21} \right) \right. \\ &- \left. \frac{n^*_{a12} + n^*_{a21}}{n_{a12} n_{a21}} (\hat{Q}_{a1} + \hat{Q}_{a2}) \right) \\ &= \frac{1}{4} \sum_{a=1}^{A} \frac{1}{n_a(n_a-1)} n_{a1} n_{a2} \left(2 \frac{n^*_{a12}}{n_{a1} n_{a2}} \hat{C}^*_{a12} - \frac{n^*_{a12}}{n_{a12} n_{a21}} (\hat{Q}_{a1} + \hat{Q}_{a2}) \right) \\ &+ \frac{1}{4} \sum_{a=1}^{A} \frac{1}{n_a(n_a-1)} n_{a1} n_{a2} \left(2 \frac{n^*_{a21}}{n_{a1} n_{a2}} \hat{C}^*_{a21} - \frac{n^*_{a21}}{n_{a12} n_{a21}} (\hat{Q}_{a1} + \hat{Q}_{a2}) \right) \\ &= \frac{1}{4} \sum_{a=1}^{A} \frac{1}{n_a(n_a-1)} n^*_{a12} \left(2\hat{C}^*_{a12} - \hat{Q}_{a1} - \hat{Q}_{a2} \right) \end{aligned}$$

$$
\begin{aligned}
+ \quad & \frac{1}{4}\sum_{a=1}^{A}\frac{1}{n_a(n_a-1)}n^*_{a21}\left(2\hat{C}^*_{a21}-\hat{Q}_{a1}-\hat{Q}_{a2}\right) \\
= \quad & \frac{1}{4}\sum_{a=1}^{A}(SQE^*_a+SQE^{**}_a) \\
= \quad & SQE^*+SQE^{**}. \qquad (4.29)
\end{aligned}
$$

De tal forma que, as hipóteses do novo teste possam ser construídas da seguinte maneira

$$
\begin{aligned}
H_0 \quad &: \quad C^*_{a12}-Q_{a1}-Q_{a2}=0 \; e \; C^*_{a21}-Q_{a1}-Q_2=0 \\
H_a \quad &: \quad C^*_{a12}-Q_{a1}-Q_{a2}\neq 0 \; e/ou \; C^*_{a21}-Q_{a1}-Q_{a2}\neq 0.
\end{aligned}
$$

para todo $a=1,2,...,A$

Este teste avalia se a ***probabilidade de um aluno com pior desempenho no vestibular ter notas melhores na graduação comparado a um aluno com melhor desempenho no vestibular*** é estatisticamente não nula. Entretanto, o objetivo da análise é avaliar se alunos de um grupo g tem melhor desempenho do que alunos pertencentes a um grupo g', de outra forma, se $C^*_{gg'}$ é maior do que $C^*_{g'g}$, e vice e versa.

Uma vez identificado que há diferença entre os grupos, para saber qual grupo obteve maior desempenho irá se olhar para a diferença entre $C^*_{gg'}$ e $C^*_{g'g}$ e propor o seguinte teste de hipóteses

$$
\begin{aligned}
H_0 \quad &: \quad C^*_{gg'}-C^*_{g'g}=0 \\
H_a \quad &: \quad C^*_{gg'}-C^*_{g'g}=0;
\end{aligned}
$$

Chama-se $DIF=\hat{C}^*_{gg'}-\hat{C}^*_{g'g}$ e a partir de métodos de re-amostragem estima-se sua variância e calcula-se um intervalo de confiança.

Para esta análise será feita apenas a abordagem não paramétrica, utilizando o método de re-amostragem por *jackknife* para estimar as variâncias e calcular os intervalos de confiança.

Capítulo 5

Aplicações

Neste capítulo serão apresentadas as aplicações com dados reais para cada uma das metodologias apresentadas no Capítulo 5. Para a metodologia 1, baseada no ganho relativo, são feitas duas abordagens, uma paramétrica (utilizando as distribuição triangular e normal e método de estimação por verossimilhança) e uma abordagem não paramétrica (utilizando técnicas de re-amostragem). Para a segunda metodologia é feito apenas a abordagem não paramétrica.

Para estimar as quantidades Q, C e calcular SQE e sua variância em cada um dos problemas propostos, foram desenvolvidas rotinas e utilizado o software R versão 2.1.1. Estas rotinas são apresentadas nos apêndices B e C.

5.1 Descrição do conjunto de dados

O conjunto de dados foi fornecido pela COMVEST (Comissão Permanente de Vestibulares) e pela DAC (Diretoria Acadêmica da Unicamp) da Universidade Estadual de Campinas. Contém informações de 7515 alunos ingressantes nos anos de 1997, 1998, 1999 e 2000, em todos os cursos de graduação oferecidos (foram excluídos os alunos ingressantes nos cursos tecnológicos ministrados no campus de Limeira). Esses alunos correspodem a uma amostra do universo de todos os alunos já ingressantes na UNICAMP, ou seja, a

população de interesse são os alunos ingressantes.

Os dados foram atualizados no final do primeiro semestre de 2007 e as informações disponíveis se referem as respostas do questionário sócio-cultural aplicado pela COMVEST, no momento da inscrição no vestibular, e informações acadêmicas tais como: nota obtida no vestibular, coeficiente de rendimento final do aluno, situação acadêmica, notas obtidas nas disciplinas cursadas.

A amostra é composta, em sua maioria, por estudantes com idades entre 16 e 24 anos (apenas 7,3% ingressantes declararam ter mais de 24 anos) de ambos os sexos, sendo 4403 (59,1%) do sexo masculino e 3051 (40,9%) do sexo feminino (faltou informação sobre o sexo para 61 alunos).

A situação acadêmica desses alunos, como pode ser vista na Tabela 5.1, foi classificada de três formas: graduados (alunos que já haviam concluído o curso), ativos (alunos que não haviam concluído o curso e que ainda estavam matriculados) e outros (em geral alunos jubilados ou desistentes). A maioria dos alunos (76,1%) já havia se graduado e 22,4% foram jubilados ou desistiram do curso, apenas 1,5% ainda estavam ativos. Há um aumento do número de ingressantes ao longo dos anos devido a maior oferta de vagas pela universidade. Tomando como base o anos de 1997, em 1998 houve um aumento no ingresso do alunos em 23,4% (1729/1393 - 1), em 1999 57,4% (2192/1393 - 1) e em 2000 o número de alunos era 58,6% maior do que em 1997 (2210/1393 - 1). O maior aumento ocorreu de 1998 para 1999, onde o crescimento foi de 24,7% (1720/2192 - 1).

Os alunos são ingressantes em 45 cursos pertencentes a todas as áreas do conhecimento. Os cursos, segundo a área, são os seguintes

1. *Exatas*: matemática, estatística, física, química, matemática aplicada, licenciatura em matemática, licenciatura em física, ciência da computação, química tecnológica, ciências da terra, geologia e geografia;

2. *Tecnológicas*: arquitetura e os cursos de engenharia agrícola, química, mecânica, elétrica, civil, de alimentos, de computação e de controle e automação;

Tabela 5.1: Distribuição da amostra total por ano segundo a situação acadêmica do aluno.

Situação	**ano de ingresso**								**Total**	
acadêmica	**1997**		**1998**		**1999**		**2000**			
	n	%	n	%	n	%	n	%	n	%
graduado	1099	78,9	1315	76,5	1641	74,9	1664	75,3	5719	76,1
ativo	5	0,4	15	0,9	39	1,8	52	2,4	111	1,5
outros	289	20,8	390	22,7	512	23,4	494	22,4	1685	22,4
Total	1393	100,0	1720	100,0	2192	100,0	2210	100,0	7515	100,0

3. *Biológicas*: biologia licenciatura e bacharelado, odontologia, medicina, enfermagem e educação física;

4. *Humanas*: letras, ciências sociais, ciências econômicas, lingüística, história, pedagogia, filosofia e licenciatura em química/física;

5. *Artes*: música, dança, artes visuais e artes cênicas.

A Tabela 5.2 apresenta a distribuição da amostra por ano, segundo a área de interesse. Como se observa, as áreas mais procuradas são as Tecnológicas e Exatas com cerca de 30% e 22% dos alunos, respectivamente, seguida da Biológica e Humanas com cerca de 20% do ingressantes cada uma. A área de artes é responsável por apenas 6,4% da amostra avaliada.

A principal característica de interesse é a origem dos alunos com relação ao tipo de escola que cursaram no 2° grau, isto é, se vieram de escolas públicas ou privadas. No questionário sócio-cultural da COMVEST a questão era feita da seguinte forma *"Qual o tipo de estabelecimento que cursou no ensino médio (2° grau)"* e apresentava as seguintes alternativas

1. somente particular;

2. somente público;

Tabela 5.2: Distribuição da amostra total por ano segundo a área.

Área	**ano de ingresso**								**Total**	
	1997		**1998**		**1999**		**2000**			
	n	%	n	%	n	%	n	%	n	%
exatas	235	16,9	416	24,2	516	23,5	515	23,3	1682	22,4
biológicas	322	23,1	333	19,4	407	18,6	415	18,8	1477	19,7
humanas	238	17,1	345	20,1	492	22,5	484	21,9	1559	20,8
artes	99	7,1	113	6,6	134	6,1	134	6,1	480	6,4
tecnológicas	499	35,8	513	29,8	643	29,3	662	30,0	2317	30,8
Total	1393	100,0	1720	100,0	2192	100,0	2210	100,0	7515	100,0

3. mais público;

4. mais particular;

5. igual tempo em ambas;

6. nenhuma das alternativas.

A partir desta questão foram considerados oriundos de escolas privadas alunos que declararam ter estudado todo o ensino médio ou a maior parte do tempo em escolas particulares. Analogamente, alunos que estudaram todo o ensino médio ou a maior parte dele em estabelecimentos públicos, foram considerados como oriundos de escolas públicas. Os alunos que assinalaram uma das duas últimas opções foram considerados como falta de informação. Houve falta de informação para 117 alunos.

Como é mostrado na Tabela 5.3, em geral, 30% dos alunos que ingressam entre 1997 e 2000 são oriundos de escolas públicas, um contigente mais de 2 vezes menor do que o de alunos que cursaran escolas privadas. Essa proporção é valida também quando observado a distribuição por sexo, cerca de 30% mulheres e 30% dos homens que ingressam na universidade estudaram em escolas públicas no segundo grau (ver Figura 5.1).

Tabela 5.3: Distribuição da amostra total por ano segundo o tipo de escola cursada no 2° grau.

Tipo Escola	**ano de ingresso**								**Total**	
do 2° grau	**1997**		**1998**		**1999**		**2000**			
	n	%	n	%	n	%	n	%	n	%
privada	965	70,5	1181	69,7	1522	70,6	1500	69,0	5168	69,9
pública	404	29,5	513	30,3	635	29,4	678	31,0	2230	30,1
Total	1369	100,0	1694	100,0	2157	100,0	2178	100,0	7398	100,0

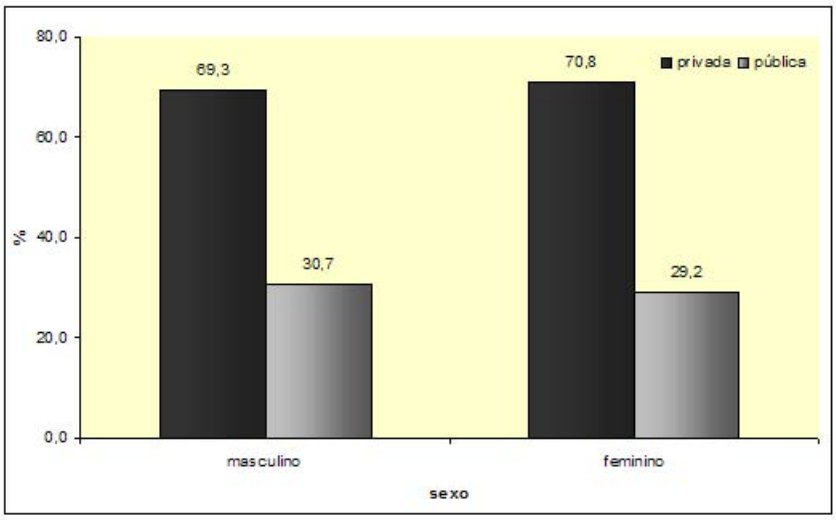

Figura 5.1: Percentagem de alunos segundo tipo de escola do segundo grau por sexo.

Com relação a área, as Exatas apresentam maior percentagem de alunos oriundos de escolas públicas com 42,6%, seguida das Humanas com 34,8% e a área com menor percentagem é a Tecnológicas com 22,1%, como se observa na Figura 5.2.

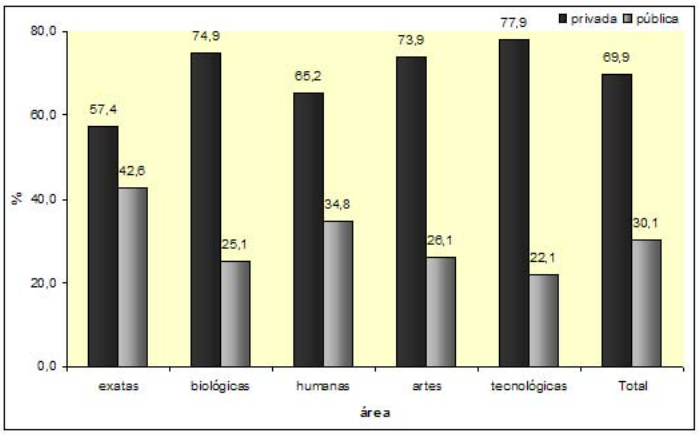

Figura 5.2: percentagem de alunos segundo tipo de escola do 2° grau por área.

Outra característica avaliada foi se o aluno "trabalhava" ou não ao entrar na universidade. Na amostra total pouco menos de um terço, 27,1% dos estudantes (1990 indivíduos), declararam que trabalhavam ao ingressar na universidade. Quando se faz essa mesma análise separadamente para cada área (Figura 5.3 observa-se que a área com maior percentagem com alunos que trabalhavam é a artes com 50,6%, seguida das Exatas com 37,5% e Humanas com 34,5%. Nas áreas Tecnológicas e Biológicas apenas 16% dos alunos trabalhavam.

Entre os sexos, a percentagem de rapazes que trabalhava é maior que a de mulheres, 29,4% contra 23,7%. Já quando se compara alunos de escola públicas e privadas, 48,0% do alunos oriundos de escolas públicas declararam que trabalhavam contra apenas 18,1% dos aluns que estudaram o ensino médio em escolas particulares (ver Tabela 5.4).

Com relação ao perfil econômico dos alunos, a única variável que se dispunha era a renda mensal familiar do aluno em salários mínimos (SM). A amostra total está distribuída da seguinte maneira, 48,2% dos estudantes declararam que a renda familiar mensal era superior a 20 SM, 29,1% renda maior que 10 e menor que 20 SM, 20,8% mais do que 3 e

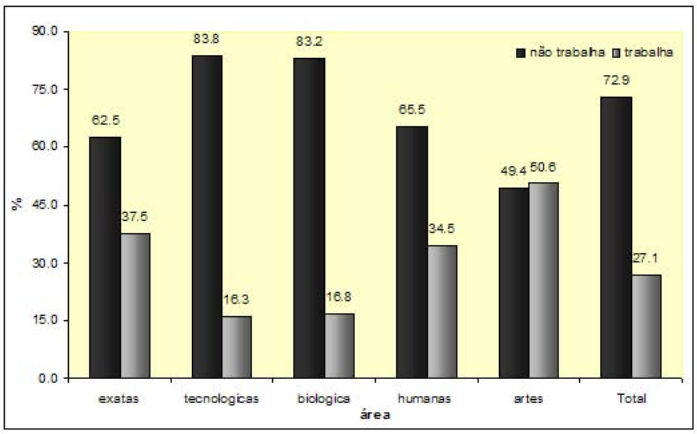

Figura 5.3: Percentagem de alunos que trabalhavam ou não ao entrar na universidade por área.

Tabela 5.4: Distribuição da amostra segundo a variável *trabalha*, por sexo e por tipo de escola do 2° grau, respectivamente.

Trabalhava ao entrar	**sexo**		**tipo de escola**		**Total**
na universidade	*masculino*	*feminino*	*particular*	*pública*	
não	70,6	76,3	81,9	52,1	72,9
sim	29,4	23,7	18,1	47,9	27,1
Total	100,0	100,0	100,0	100,0	100,0

menos do que 10 SM e apenas 2% declaram ter renda menor do que 3 SM.

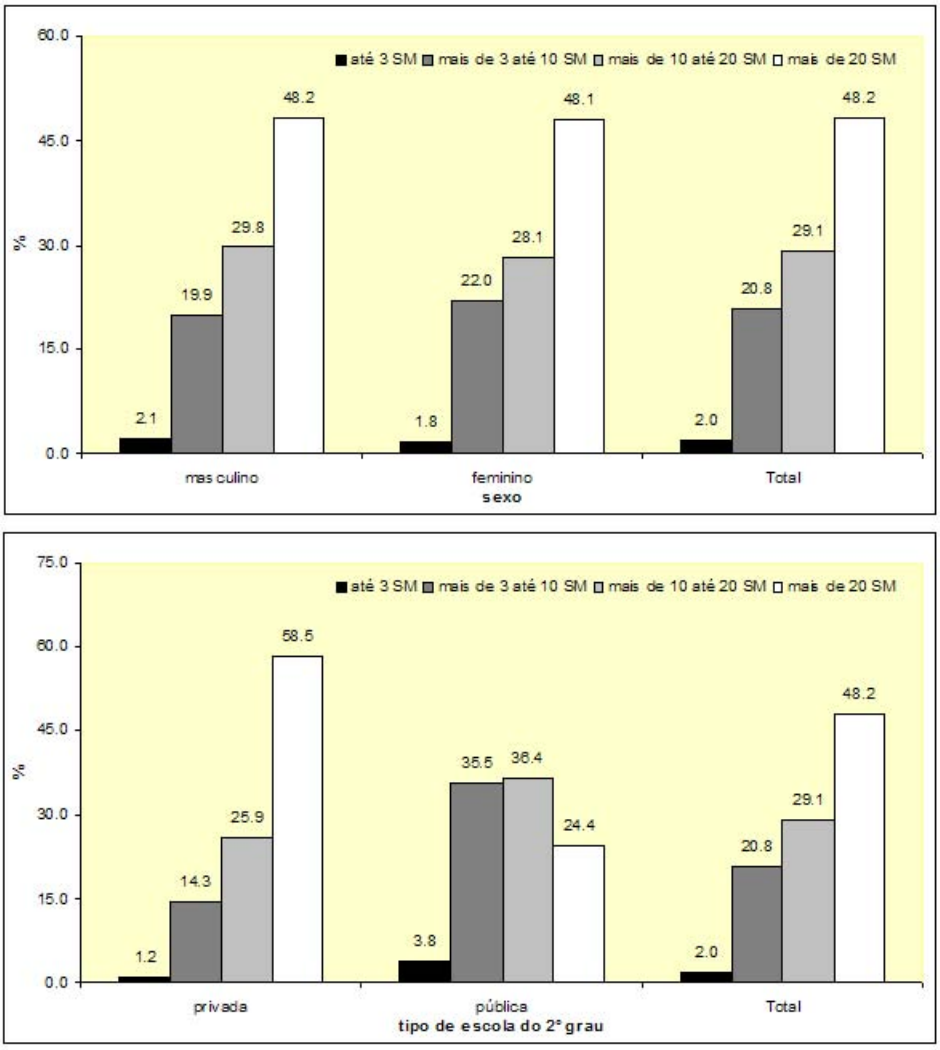

Figura 5.4: Distribuição da amostra segundo a renda familiar mensal, por sexo e por tipo de escola do 2° respectivamente.

A Figura 5.4 apresenta o gráfico com a distribuição dos estudando segunda a renda por sexo e por tipo de escola do 2° grau, respectivamente. Não há grande diferenças na distribuição da renda mensal familiar entre os sexos. Entretanto, quando se compara alunos oriundos de escolas particulares com alunos oriundos de escolas públicas, nota-se que os primeiros apresentam rendas maiores, 58,5% destes alunos declararam renda acima

de 20 SM e 25,9% renda familiar mensal entre 10 e 20 SM, ou seja, 84,4% dos estudantes oriundos de escola particulares declaram ter renda familiar mensal acima de 10 SM. Entre os alunos que estudaram em escola públicas, 35,5% declaram renda familiar entre 3 e 10 SM, 36,4% renda entre 10 e 20 SM e apenas 24,4% renda familiar mensal superior a 20 SM. Comparado com o grupo anterior (alunos de escola privadas) 60,8% declararam renda acima 10 SM (23,6% a menos).

O ganho relativo, ou simplesmente ganho, já foi apresentado e cuidadosamente analisado no Capítulo 1. A Tabela 5.5 apresenta algumas medidas resumo para o ganho relativo segundo alguns grupos formados de acordo com o sexo e o tipo de escola e se o aluno trabalhava ao entrar na universidade.

Tabela 5.5: Medidas resumo para o ganho relativo segundo o grupo de interesse.

Grupos	n	média	D.P.	mínimo	máximo
tipo de escola					
particular	5156	-0,0195	0,3476	-0,984	0,984
pública	2223	0,0465	0,3632	-0,976	0,958
sexo					
masculino	4356	-0,0405	0,3544	-0,984	0,984
feminino	3023	0,0593	0,3442	-0,947	0,957
trabalhava					
não	5414	0,0018	0,3474	-0,984	0,968
sim	2029	-0,0030	0,3691	-0,960	0,984
interação sexo e tipo de escola do 2º grau					
masculino particular	3017	-0,0588	0,3478	-0,984	0,984
masculino pública	1339	0,0006	0,3655	-0,976	0,958
feminino particular	2139	0,0359	0,3397	-0,927	0,947
feminino pública	884	0,1160	0,3485	-0,947	0,957

Os alunos oriundos de escola pública apresentaram ganho médio (0,0465) maior comparado aos alunos que estudaram escolas particulares (-0,0195). Quando comparado os sexos, as mulheres apresentaram ganho médio igual a 0,0593 contra -0,0405 dos homens.

Avaliando a interação entre o sexo e o tipo de escola, as disparidades entre tipo de escola são maiores entre as mulheres do que entre os homens. A diferença no ganho relativo médio entre alunos de escolas públicas e alunos de escola privadas é de 0,0804 entre as mulheres e 0,0594 entre os homens. Essas divergências também podem ser observadas a partir dos gráficos de caixas apresebtados na Figura 5.5.

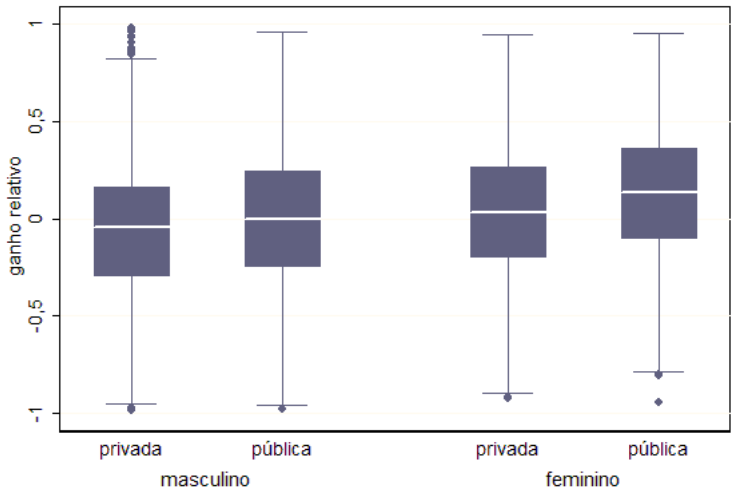

Figura 5.5: Gráfico de Caixas para o ganho relativo segundo o sexo e o tipo de escola do 2º grau.

Um segundo conjunto de dados, fornecido pela Diretoria Acadêmica da UNICAMP (DAC), contém as notas de todas as disciplinas cursadas pelos alunos que ingressaram entre 1997 e 2000, de todos os cursos de graduação. São oferecidas da UNICAMP mais de 3000 disciplinas para os 45 cursos oferecidos.

O número de disciplinas cursadas pelos alunos varia de 1 até 136 (sem contar as repetições, isto é, disciplinas que foram cursadas mais de uma vez). Na análise de Diversidade a partir das notas obtidas pelos alunos, serão considerados apenas os alunos que cursaram mais de 20 disciplinas para que se tenha um número razoável de comparações, dado que, o que se compara na metodologia proposta na seção 5.2 são as notas das disciplinas cursadas em comum entre dois alunos. O número total de alunos que se enquadram nesta condição é 6459, sendo 1195 ingressantes em 1997, 1458 em 1998, 1874 em 1999 e 1932 em 2000.

O número de vezes que um aluno cursa uma determinada disciplina varia de 1 a 12 vezes. Como já descrito na seção 5.2, quando um aluno $\mathbb{A}$ houver cursado uma determinada disciplinas por mais vezes que um aluno $\mathbb{B}$, será considerado que o segundo obteve melhor desempenho na disciplina, independente da média com que ambos foram aprovados.

5.2 Análise da Diversidade do ganho relativo

5.2.1 Abordagem Paramétrica

Na abordagem paramétrica as variâncias da estatísticas do teste serão estimadas assumindo uma função de probabilidade para a variável ganho relativo.

Utilizando a Distribuição Normal

Na primeira abordagem feita na análise de diversidade do ganho relativo, assume-se que a variável vem de uma distribuição Normal com parâmetros μ_g e σ_g. O objetivo é avaliar se, dado dois grupos, eles são homogêneos ou se um grupo apresenta maior ganho

relativo do que o outro. Equivalentemente, é dizer que as distribuições do ganho para cada grupo, ambas normais, apresentam os mesmos parâmetros.

O primeiro passo para a análise é estimar os parâmetros da distribuição para cada grupo estudado. As variáveis avaliadas foram: tipo de escola pública do 2° grau, sexo e se trabalhava ao entrar na universidade. A Tabela 5.6 apresenta estimativas da média e variância do ganho relativo, segundo o grupo de interesse.

Tabela 5.6: Média e variância do ganho relativo segundo o grupo de interesse.

Grupos	**n**	**média**	**variância**
tipo de escola			
particular	5156	-0,0195	0,1208
pública	2223	0,0465	0,1319
sexo			
masculino	4356	-0,0405	0,1256
feminino	3023	0,0593	0,1184
trabalhava			
não	5414	0,0018	0,1207
sim	2029	-0,0030	0,1362

A Figura 5.6 apresenta a curva da distribuição normal para cada grupo de interesse segundo a característica avaliada. Com relação ao tipo de escola do 2° grau, os estudantes de escola públicas apresentaram estimativa de média maior e menor estimativa de variância, comparados aos alunos de escola particulares. Entre os sexos, as mulheres apresentaram estimativas de média e variância superiores as estimativas para o ganho relativo dos homens. Com relação ao fator trabalho, ambos os grupos apresentaram médias próximas, entretanto a variância dos alunos que trabalhavam é pouco menor do que a estimativa da variância dos que não trabalhavam.

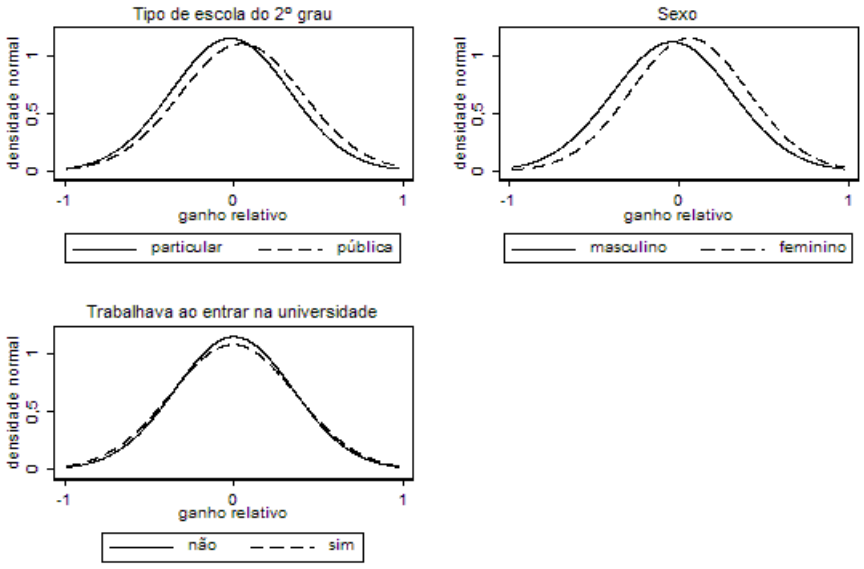

Figura 5.6: Curva normal estimada para o ganho relativo segundo o grupo de interesse.

O passo seguinte foi obter as estimativas dos momentos de ordem 1,2,3 e 4 da distribuição para cada um dos grupos. Para isso se usou o Lema 4.1 e as estimativas são apresentadas na Tabela 5.7.

E então foi feita a análise de diversidade estimando as quantidades Q_g e C_{gg} e calculando SQE. A variância de SQE foi calculada a partir das estimativas da Tabela 5.7 e obtido um intervalo de 95% de confiança conforme descrito em (4.12).

Como se observa na Tabela 5.8, apenas o intervalo de confiança para o fator sexo não compreende o valor 0, portanto, só houve diferença significativa, ao nível de 5%, para a diversidade entre os homens e mulheres. E como apresentada anteriormente, as mulheres tiveram maior ganho relativo médio. Em média os estudantes do sexo masculino estão perdendo postos (média do granho relativo = -0,0405) enquanto as mulheres estão, em média, ganhando postos (ganho relativo médio = 0,0593).

Tabela 5.7: Estimativas dos momentos da distribuição normal segundo o grupo de interesse.

Grupos	$\hat{\mu}$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$
tipo de escola				
particular	-0,0195	0,1212	-0,0071	0,0441
pública	0,0465	0,1341	0,0185	0,0539
sexo				
masculino	-0,0405	0,1272	-0,0153	0,0485
feminino	0,0593	0,1220	0,0213	0,0446
trabalhava				
não	0,0018	0,1207	0,0006	0,0437
sim	-0,0030	0,1362	-0,0012	0,0556

Tabela 5.8: Análise da Diversidade - utilizando a distribuição normal.

Grupos	$\hat{Q}_g$	$\hat{C}_{gg'}$	SQE	$d.p.(SQE)$	**IC95%**	
tipo de escola						
particular	0,2416					
pública	0,2638	0,2604	0,0032	0,0028	-0,0022	0,0086
sexo						
masculino	0,2511					
feminino	0,2369	0,2590	0,0072	0,0028	0,0017	0,0127*
trabalhava						
não	0,2416					
sim	0,2735	0,2569	-0,0003	0,0028	-0,0057	0,0052

* fator significativo ao nível de 5%

Utilizando a Distribuição Triangular

Nesta seção irá se assumir que o ganho relativo vem de uma distribuição triangular. Os fatores avaliados foram os mesmos para o caso em que é considerada a distribuição normal (tipo de escola do 2° grau, sexo e trabalha). O primeiro passo é estimar os parâmetros da distribuição triangular para cada um dos grupos avaliados. Para isso, como apresentado na seção 6.2.1, foi aplicado o método de máxima verossimilhança e utilizado o software MLE Estimator. A Tabela 5.9 apresenta as estimativas dos parâmetros.

Tabela 5.9: Estimativas dos parâmetros da distribuição Triangular segundo o grupo de interesse.

Grupos	$\hat{a}$	$\hat{m}$	$\hat{b}$
tipo de escola			
particular	-0,9863	-0,0145	0,9850
pública	-0,9795	0,1111	0,9681
sexo			
masculino	-0,9890	-0,0444	0,9853
feminino	-0,9487	0,1091	0,9596
trabalhava			
não	-0,9858	0,0000	0,9700
sim	-0,9729	0,0000	0,9880

A Figura 5.7 apresenta a curva da função densidade da triangular estimada para cada um dos grupos de interesse. Quando comparados alunos que estudaram em escola públicas ou particulares no ensino médio, a moda estimada para o primeiro é 0,1111 e para o segundo -0,0145. Já com relação ao sexo, as mulheres apresentaram moda estimada igual a 0,1091 e os homens -0,0444. A distribuição para alunos que trabalhavam é muito próxima da distribuição dos alunos que não trabalhavam.

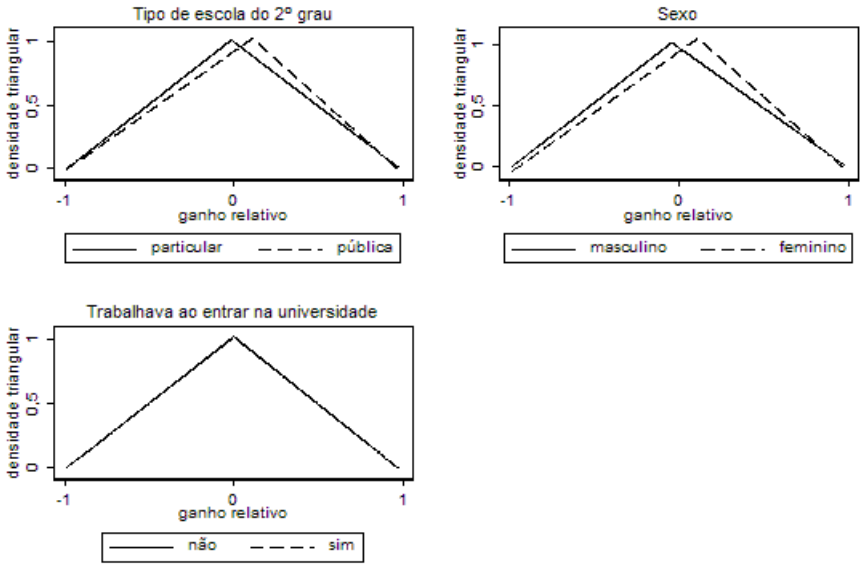

Figura 5.7: Curva triangular estimada para o ganho relativo segundo o grupo de interesse.

A próxima etapa é calcular as estimativas dos 4 primeiros momentos da distribuição triangular para cada um dos grupos de interesses. Os resultados são apresentados na Tabela 5.10.

Os resultados da análise de diversidade são mostrados na Tabela 5.11. As estimativas para a variância de SQE são maiores do que as observadas na análise utilizando a distribuição normal, como já era esperado. Mesmo com maiores estimativas da variância o fator sexo ainda foi significativo ao nível de 5%, quando se assumiu a distribuição triangular.

Tabela 5.10: Estimativas dos momentos da distribuição triangular segundo o grupo de interesse.

Grupos	$\hat{\mu}$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\sigma}^2$
tipo de escola					
particular	-0,0053	0,1620	-0,0017	0,0629	0,1619
pública	0,0332	0,1599	0,0085	0,0606	0,1588
sexo					
masculino	-0,0160	0,1628	-0,0050	0,0635	0,1625
feminino	0,0067	0,1365	0,0025	0,0497	0,1364
trabalhava					
não	-0,0053	0,1594	-0,0030	0,0610	0,1594
sim	0,0050	0,1602	0,0029	0,0616	0,1602

Tabela 5.11: Análise da Diversidade - utilizando a distribuição triangular.

Grupos	$\hat{Q}$	$\hat{C}$	SQE	$d.p.(SQE)$	**IC95%**	
tipo de escola						
particular	0,2416					
pública	0,2638	0,2604	0,0032	0,0035	-0,0036	0,0101
sexo						
masculino	0,2511					
feminino	0,2369	0,2590	0,0072	0,0034	0,0005	0,0140*
trabalhava						
não	0,2416					
sim	0,2735	0,2569	-0,0003	0,0035	-0,0070	0,0065

* fator significativo ao nível de 5%

5.2.2 Abordagem Não Paramétrica

Nesta abordagem as estimativas da variância foram calculadas a partir do método de re-amostragem de *jackknife* descrito na seção 4.1.2. Também foi considerado o problema de multifatores na análise de diversidade apresentado no capítulo 4, este problema foi considerado ao se avaliar a interação entre dois fatores de interesse (tipo de escola do 2° grau e sexo).

A Tabela 5.12 apresenta os valores estimados da SQE (Soma de Quadrados Entre grupos), o desvio padrão e o Intervalo de 95% de Confiança obtido a partir do *jackknife*. Os IC95% foram construidos utilizando a seguinte formula: $SQE \pm 1,96 DPjack$.

Tabela 5.12: Análise da diversidade a partir da abordagem não paramétrica.

grupos	**SQE**	**D.P.jack**	**IC95%**	
escola	0.00324	0,0007011	0,0019	0,0046*
sexo	0.00278	0,0010151	0,0008	0,0048*
trabalha	-0,00025	0,0005017	-0,0012	0,0007
escola x sexo	0.01039	0,0012890	0,0079	0,0129*
esola|sexo	0.00761	0,001634	0,0044	0,0108*
sexo|escola	0.00715	0,001462	0,0043	0,0100*

* fator significativo ao nível de 5%

A variância estimada pelo método de re-amostragem é menor comparada as estimativas obtidas ao se assumir uma distribuição para o ganho relativo (caso paramétrico). Aqui o único fator que não foi significativo ao nível de 5% foi o fator "trabalhava ao entrar na universidades".

A interação entre sexo e tipo de escola foi significativa, ao nível de 5%. Como pode ser observado na Figura 5.5, a diferença no ganho relativo entre alunos oriundos de escola públicas e privadas, quando se compara apenas estudantes do sexo feminino é maior do que quando comparados apenas os estudantes do sexo masculino.

Ao se avaliar o fator tipo escola sem a influência do sexo (escola dado sexo) e o fator sexo sem a influência da escola (sexo dado escola), ambos os fatores foram significativos.

A Figura 5.8 mostra o histograma das estimativas da SQE obtidas pelo método de re-amostragem para cada um dos grupos de interesse.

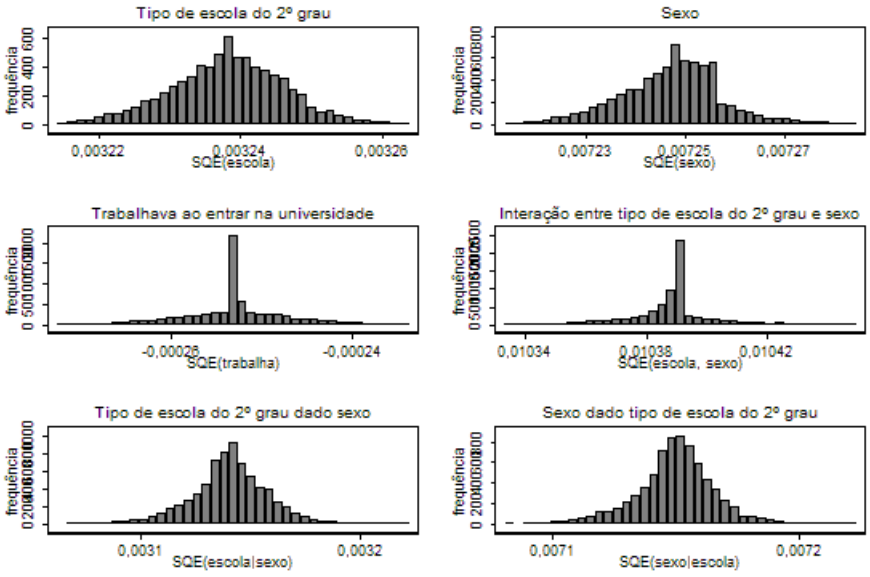

Figura 5.8: Histograma para o *jackknife* da SQE.

5.3 Análise da Diversidade a partir das notas das disciplinas

Na análise de diversidade a partir das disciplinas cursadas durante a graduação, foram avaliados apenas dois fatores: tipo de escola do 2° grau e sexo. A Tabela 5.13 apresenta o resultado das análises para cada um dos fatores, respectivamente.

Os dois fatores foram significativos, ao nível de 5%, pois nenhum dos intervalos de confiança calculado compreendem o valor 0. Isto é valido tando na avaliação de cada ano individualmente, quanto na análise da amostra total.

Para determinar quais os grupos que apresentaram maior desempenho, olha-se para as quantidades $\hat{C}^*_{12}$ e $\hat{C}^*_{21}$. O intervalo de confiança para a diferença entre elas é apresentado na Tabela 5.14.

Tabela 5.13: Análise da diversidade para as notas das disciplinas - avaliando o tipo de escola do 2° grau.

$grupo_1$	n_1	$\hat{Q}_1$	$\hat{C}^*_{12}$	SQE^*	$\hat{d.p._{jack}}(SQE^*)$	**IC95%**	
$grupo_2$	n_1	$\hat{Q}_2$	$\hat{C}^*_{21}$	SQE^{**}	$\hat{d.p._{jack}}(SQE^{**})$	**IC95%**	
particular 00	1323	0,1354	0,0963	-0,0172	0,0013	-0,0197	-0,0147*
pública 00	609	0,1828	0,0583	-0,0160	0,0012	-0,0183	-0,0136*
particular 99	1329	0,1359	0,0903	-0,0179	0,0013	-0,0205	-0,0154*
pública 99	545	0,1749	0,0576	-0,0134	0,0016	-0,0165	-0,0103*
particular 98	1025	0,1438	0,1000	-0,0146	0,0014	-0,0174	-0,0119*
pública 98	433	0,1575	0,0463	-0,0135	0,0013	-0,0160	-0,0110*
particular 97	843	0,1276	0,0901	-0,0168	0,0017	-0,0201	-0,0135*
pública 97	352	0,1763	0,0566	-0,0138	0,0013	-0,0164	-0,0112*
Total part.	4520	0,1357	0,0942	-0,0166	0,0014	-0,0194	-0,0139*
Total públ.	1939	0,1729	0,0547	-0,0142	0,0013	-0,0168	-0,0115*
masculino 00	1075	0,1803	0,0896	-0,0189	0,0014	-0,0216	-0,0161*
feminino 00	857	0,1278	0,0471	-0,0215	0,0013	-0,0240	-0,0190*
masculino 99	1102	0,1815	0,0809	-0,0197	0,0018	-0,0233	-0,0160*
feminino 99	772	0,1150	0,0467	-0,0196	0,0020	-0,0234	-0,0158*
masculino 98	829	0,1790	0,0840	-0,0194	0,0020	-0,0233	-0,0154*
feminino 98	629	0,1225	0,0492	-0,0204	0,0023	-0,0249	-0,0160*
masculino 97	689	0,1821	0,0809	-0,0170	0,0023	-0,0215	-0,0125*
feminino 97	506	0,1033	0,0434	-0,0212	0,0023	-0,0256	-0,0168*
Total masc.	3695	0,1807	0,0838	-0,0187	0,0019	-0,0224	-0,0150*
Total fem.	2764	0,1172	0,0466	-0,0207	0,0019	-0,0245	-0,0169*

* fator significativo ao nível de 5%

Como se observa no caso da variável ganho relativo, $\hat{C}^*_{12} > \hat{C}^*_{21}$, portanto pode-se concluir que a $\mathbf{P}$*(de um aluno de escola pública, que teve desempenho no vestibular inferior ao de um aluno de escola particular, ter notas melhores durante a graduação)* é maior que a $\mathbf{P}$*(de um aluno de escola particular, que teve desempenho no vestibular inferior ao de um aluno de escola pública, ter notas melhores durante a graduação).* A estimativa de C^*_{12} é 0,0942 e de C^*_{21} é 0,0547, para toda a amostra.

Analogamente, $\mathbf{P}$*(de um aluno do sexo feminino, que teve desempenho no vestibular inferior ao de um aluno do sexo masculino, ter notas melhores durante a graduação)* é maior que a $\mathbf{P}$*(de um aluno do sexo masculino, que teve desempenho no vestibular inferior ao de um aluno do sexo feminino, ter notas melhores durante a graduação).* As estimativas dessas proporções, para a amostra total, são $C^*_{12} = 0{,}0838$ e C^*_{21} 0,0466.

Tabela 5.14: Análise da diferença entre $\hat{C}^*_{12}$ e $\hat{C}^*_{21}$.

grupos	$\hat{C}^*_{12} - \hat{C}^*_{21}$	$\hat{dp}_{JACK}$	**IC95%**	
Escola				
2000	0,0380	0,0057	0,0269	0,0491*
1999	0,0328	0,0056	0,0217	0,0438*
1998	0,0537	0,0064	0,0411	0,0663*
1997	0,0335	0,0070	0,0198	0,0472*
Total	0,0395	0,0062	0,0274	0,0516*
Sexo				
2000	0,0425	0,0048	0,0331	0,0519*
1999	0,0342	0,0046	0,0253	0,0432*
1998	0,0348	0,0053	0,0243	0,0452*
1997	0,0375	0,0065	0,0248	0,0502*
Total	0,0372	0,0053	0,0269	0,0476*

* fator significativo ao nível de 5%

As Figuras 5.9, 5.10, 5.11 e 5.12 apresentam os histogramas das replicações *jackknife*

das estatítsicas SQE^{*}, SQE^{**}, e a diferença $\hat{C}^{*}_{12} - \hat{C}^{*}_{21}$ por ano de ingresso.

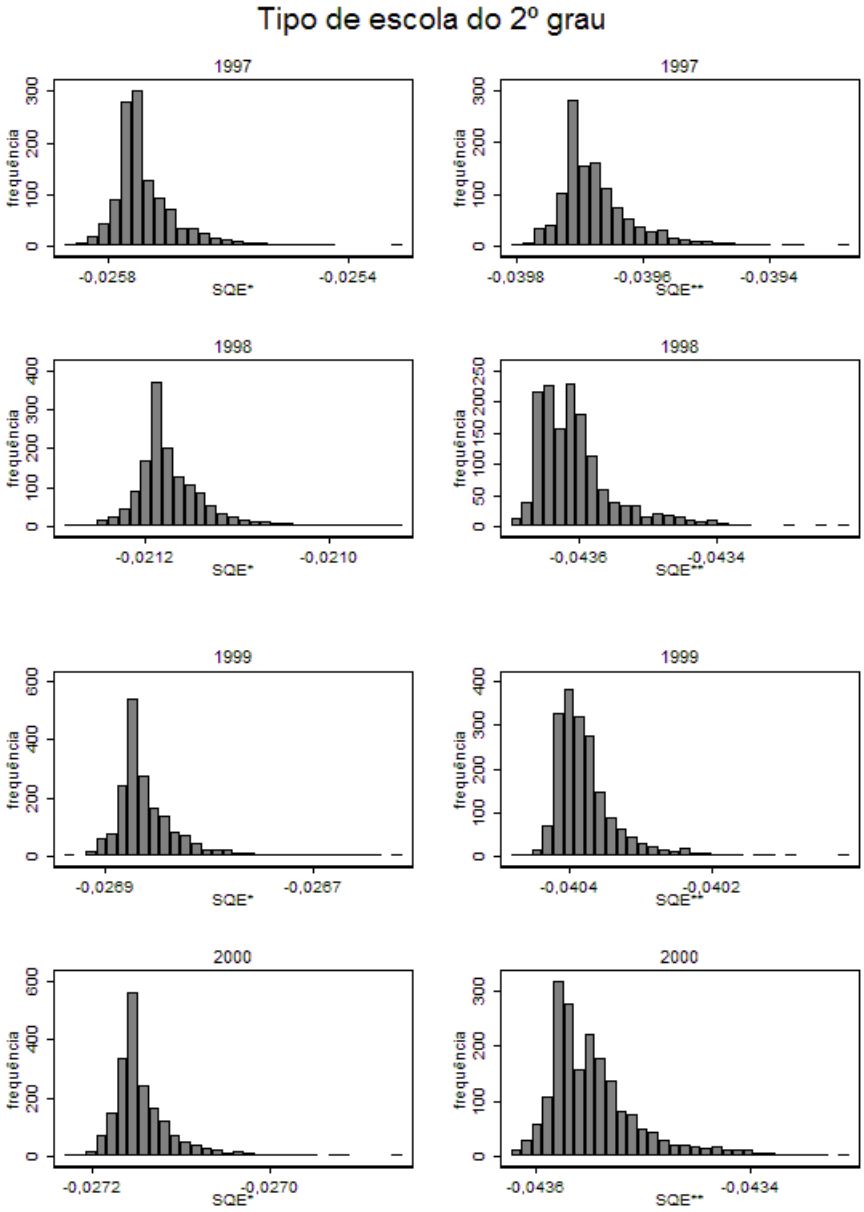

Figura 5.9: Histograma para o *jackknife* de SQE^* e SQE^{**}, Tipo de escola do 2° grau.

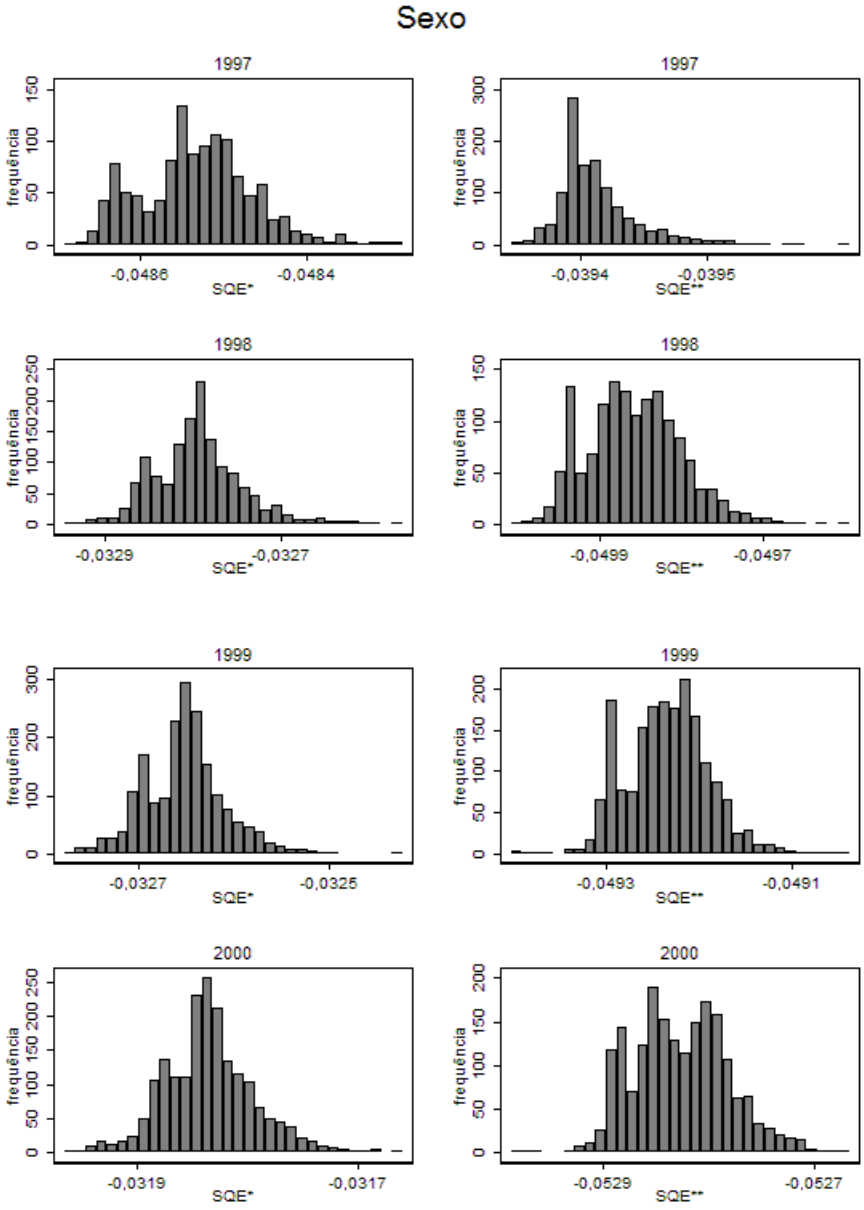

Figura 5.10: Histograma para o *jackknife* de SQE^* e SQE^{**}, Sexo.

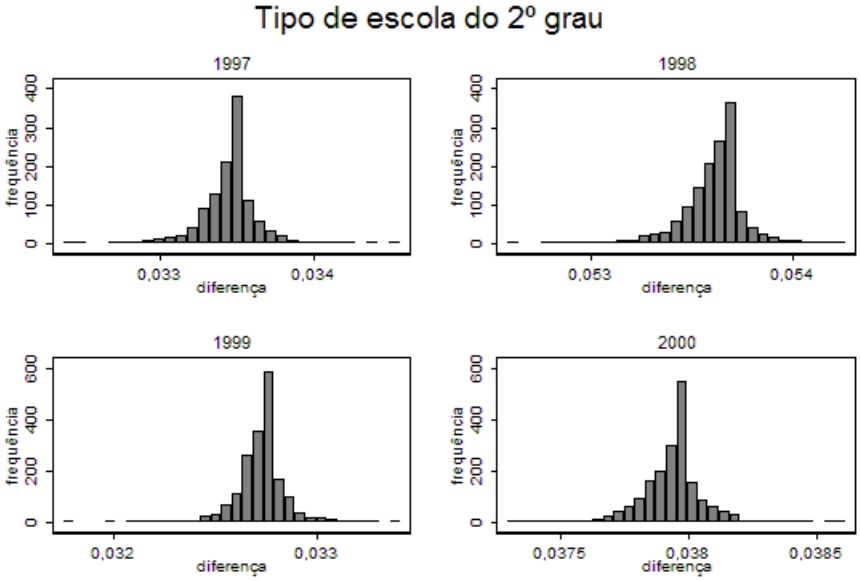

Figura 5.11: Histograma para o *jackknife* de $\hat{C}^*_{12} - \hat{C}^*_{21}$, Tipo de escola do 2° grau.

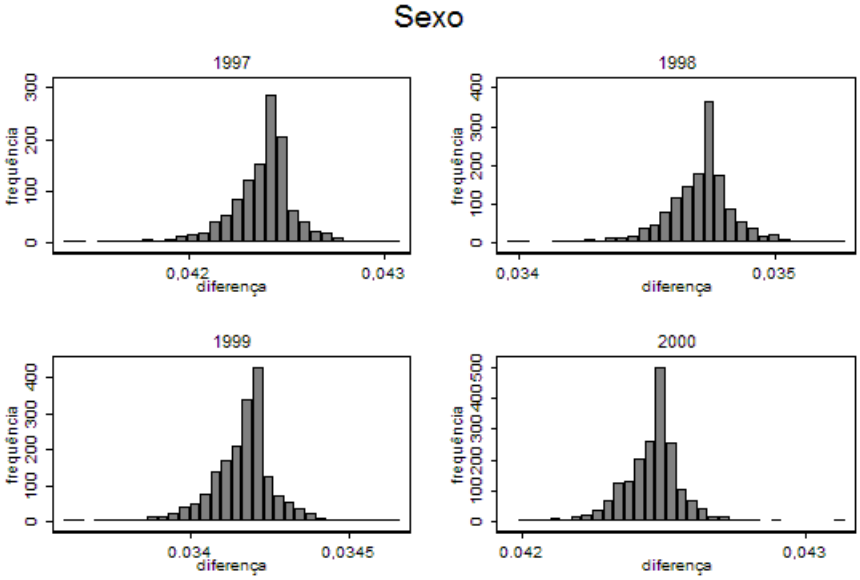

Figura 5.12: Histograma para o *jackknife* de $\hat{C}^*_{12} - \hat{C}^*_{21}$, Sexo.

5.4 Considerações finais

O objetivo do trabalho é propor novas metodologias para a análise do desempenho dos alunos da UNICAMP durante o curso de graduação. E exemplificar a metodologia a partir de dados reais, dados estes correspondentes a uma amostra do universo de alunos ingressantes na UNICAMP. Alguns pontos podem ser aperfeiçoados e ficam como propostas para novos trabalhos.

Uma das vantagens das metodologias baseadas nas medidas de diversidade propostas por Rao(1982), é que nela todos os indíviduos são comparados 2 a 2, avaliando toda a variabilidade dos dados que proporciona estatísticas menos suscetíveis a perturbações. A metodologia 2 é mais robusta do que a metodologia 1, pois utiliza as informações de todas as disciplinas cursadas pelos alunos, enquanto que na primeira é levado em consideração apenas o coeficiente de rendimento dos alunos (que é uma média das médias de todas as disciplinas cursadas, padronizada para variar entre 0 e 1) e a nota final do vestibular.

Com relação as abordagens paramétrica e não paramétrica, a segunda é mais robusta, pois não faz suposição sobre a distribuição dos dados, além de que é muito mais simples de se obter as estimativas das variâncias por técnicas de re-amostragem. Ao se assumir uma distribuição para o conjunto de dados, escolher yna distribuição que não se adequa bem a eles pode acarretar uma sobre-estimãção das variâncias. Como se vê na secção 5.2, as estimativas da variância de SQE na análise paramétrica são maiores do que ba análise não paramétrica, e quando se assume a distribuição triangular, as estimativas são ainda maiores. Entrentanto, uma vez que se tem certeza da verdadeira distribuição dos dados, os teste paramétricos são mais indicados por apresentarem testes mais poderosos.

Quando é feita a análise de diversidade do ganho relativo a partir da abordagem paramétrica, não foi feito a análise da interação entre o tipo de escola do 2° e o sexo, por exemplo. Isto porque, para estimar a variância da Soma de Quadrados Entre grupos (SQE) é preciso estimar as co-variâncias entre as estatísticas $\hat{C}_{gg'}$ e $\hat{Q}_g$ que possuem fórmulas bastante complexas, descritas na seção 3.3.2.. Fica como sugestões para trabalhos futuros analisar maneiras mais eficientes de se obter essas variâncias.

Na análise de diversidade a partir das disciplinas cursadas na graduação, as rotinas desenvolvidas têm um custo computacional muito alto, são bastante demoradas. Devido a esse fator, não foram analisadas mais variáveis e a interação entre sexo e escola. Desenvolver rotinas mais eficazes são propostas para novos trabalhos.

Apêndice A

Distribuição Triangular

Se Z tem distribuição Triangular limitada em $[a, b]$ e moda igual a m, então a função densidade de probabilidade (f.d.p) de Z, $f(z)$ é dada por

$$f(z|a,m,b) = \begin{cases} \frac{2}{b-a}\frac{z-a}{m-a} & se \quad a \leq z \leq m \\ \frac{2}{b-a}\frac{b-z}{b-m} & se \quad m \leq z \leq b \\ 0 & caso\ contrário. \end{cases} \tag{1.1}$$

O gráfico da f.d.p. da variável Z é mostrado na Figura A.1. A função de distribuição acumulada da variável Z, $F(z)$, é então

$$F(z) = P(Z \leq z) = \begin{cases} 0 & se \quad z < a \\ \frac{m-a}{b-a}(\frac{z-a}{m-a})^2 & se \quad a \leq z \leq m \\ 1 - \frac{b-m}{b-a}(\frac{b-z}{b-a})^2 & se \quad m \leq z \leq b \\ 1 & se \quad z > b. \end{cases} \tag{1.2}$$

Seja os momentos de ordem k da varíavel Z $\mu^k = EZ^k$ para $k = 1, 2, ...$ Então

$$\begin{aligned} \mu^1 &= \int_a^b z f(z) dz \\ &= \int_a^m z \frac{2}{(b-a)} \frac{z-a}{(m-a)} dz \ + \ \int_m^b z \frac{2}{(b-a)} \frac{b-z}{(b-m)} dz \\ &= c_1 \int_a^m z^2 - az dz \ + \ c_2 \int_m^b bz - z^2 dz \end{aligned}$$

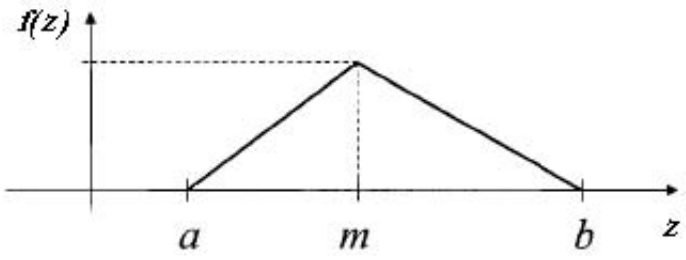

Figura A.1: Função densidade de probabilidade para uma variável aleatória Z, com distribuição triangular em $[a, b]$ e moda igual a m.

$$\begin{aligned} &= c_1 \left| \frac{z^3}{3} - \frac{az^2}{2} \right|_a^m + c_2 \left| \frac{bx^2}{2} - \frac{x^3}{3} \right|_m^b \\ &= c_1 \Big(\frac{m^3}{3} - \frac{am^2}{2} - \frac{a^3}{3} + \frac{a^3}{2} \Big) + c_2 \Big(\frac{b^3}{2} - \frac{b^3}{3} - \frac{bm^2}{2} + \frac{m^3}{3} \Big), \end{aligned}$$

em que $c_1 = \frac{2}{(b-a)} \frac{1}{(m-a)}$ e $c_2 = \frac{2}{(b-a)} \frac{1}{(b-m)}$.

$$\begin{aligned} \mu^2 &= c_1 \int_a^m z^2(z-a)dz + c_2 \int_m^b z^2 b - z dz \\ &= c_1 \left| \frac{z^4}{4} - \frac{az^3}{3} \right|_a^m + c_2 \left| \frac{bz^3}{3} - \frac{z^4}{4} \right|_m^b \\ &= c_1 \Big(\frac{m^4}{4} - \frac{am^3}{3} - \frac{a^4}{4} + \frac{a^4}{3} \Big) + c_2 \Big(\frac{b^4}{3} - \frac{b^4}{4} - \frac{bm^3}{3} + \frac{m^4}{4} \Big). \end{aligned}$$

Em geral,

$$\mu^k = c_1 \Big(\frac{m^{k+2}}{k+2} - \frac{am^{k+1}}{k+1} - \frac{a^{k+2}}{k+2} + \frac{a^{k+2}}{k+1} \Big) + c_2 \Big(\frac{b^{k+2}}{k+1} - \frac{b^{k+2}}{k+2} - \frac{bm^{k+1}}{k+1} + \frac{m^{k+2}}{k+2} \Big), \quad (1.3)$$

para todo $k = 1, 2, ...$

Para o caso em que $a = -1$, $b = 1$ e $m = 0$, ou seja, se Z tem distribuição Triangular em $[-1, 1]$ com moda igual a 0, então

$$f_Z(z) = \begin{cases} z+1 & se \quad -1 \leq z \leq 0 \\ 1-z & se \quad 0 \leq z \leq 1 \\ 0 & caso\ contrário. \end{cases} \quad (1.4)$$

$$F_Z(z) = \begin{cases} 0 & se \quad z < -1 \\ \frac{(z+1)^2}{2} & se \quad -1 \leq z \leq 0 \\ 1 - \frac{(1-z)^2}{2} & se \quad 0 \leq z \leq 1 \\ 1 & se \quad z > 1. \end{cases} \tag{1.5}$$

as constantes c_1 e c_2 serão iguais a 1 e os momentos μ^k serão

$$\mu^k \;=\; \left(-\frac{(-1)^{k+2}}{k+2} + \frac{(-1)^{k+2}}{k+1}\right) + \left(\frac{1}{k+1} - \frac{1}{k+2}\right).$$

Para $k = 1, 2, 3, 4$, obtém-se o seguinte: $\mu^1 = \mu^3 = 0$, $\mu^2 = \frac{1}{6}$ e $\mu^4 = \frac{1}{15}$.

Estimação dos parâmetros de uma distribuição Triangular

Sejam $Z_1, Z_2, ..., Z_n$ uma mostra aleatória de tamanho n de uma variável aleatória Z com distribuição triangular em $[a, b]$ e moda m. O vetor das estatísticas de ordem é dado por $\underline{\mathbf{Z}} = (Z_{(1)}, Z_{(2)}, ..., Z_{(n)})$, em que $Z_{(1)} \leq Z_{(2)} \leq ... \leq Z_{(n)}$. Utilizando a função de probabilidade descrita em 1.1, a verossimilhança para $\underline{\mathbf{Z}}$ é dada por

$$\begin{aligned} L(\underline{\mathbf{Z}}|a, m, b) &= \prod_{i=1}^{n} f(Z_{(i)}|a, m, b) \\ &= \left(\frac{2}{b-a}\right)^n \left\{ \prod_{i=1}^{r} \frac{Z_{(i)} - a}{m - a} \prod_{i=r+1}^{n} \frac{b - Z_{(i)}}{b - m} \right\}, \end{aligned} \tag{1.6}$$

em que r é implicitamente definido por $Z_{(r)} \leq m < Z_{(r+1)}$, $Z_{(0)} \equiv a$ e $Z_{(n+1)} \equiv b$.

Portanto, segue que para valores de a e b fixados, satisfazendo $a < Z_{(1)}$ e $b > Z_{(n)}$, têm-se que

$$\max_{a \leq m \leq b} L(\underline{\mathbf{Z}}|a, m, b) \;=\; \left(\frac{2}{b-a}\right)^n \{M(a, b, \hat{r}(a, b))\}, \tag{1.7}$$

em que

$$\hat{r}(a, b) = \arg \max_{r \in \{1,...,n\}} M(a, b, r) \;\; e \;\; M(a, b, r) = \prod_{i=1}^{r-1} \frac{Z_{(i)} - a}{Z_{(r)} - a} \prod_{i=r+1}^{n} \frac{b - Z_{(i)}}{b - Z_{(r)}}. \tag{1.8}$$

O estimador de máxima verossimilhança (EMV) para a moda m (como uma função de a e b) é dado por $\hat{m}(a, b) = Z_{(\hat{r}(a,b))}$. Note que a função $\hat{r}(a, b)$ indica em qual estatística

de ordem o EMV do parâmetro m é atingido como uma função dos limites inferior a e superior b.

Da equação (1.8) tem-se que

$$\max_{S(a,m,b)} \left[\log \left\{L(\underline{\mathbf{Z}}; a, m, b)\right\}\right] = \max_{a<X_{(1)}, b>X_{(n)})} \left[\log \left\{n \log 2 + G(a, b)\right\}\right], \tag{1.9}$$

em que o conjunto

$$S(a, m, b) \;=\; \left\{(a, m, b) | a < Z_{(1)}, b > Z_{(n)}, a \leq m \leq b\right\}$$

e a função

$$G(a, b) \;=\; \log \left\{M(a, b, \hat{r}(a, b))\right\} - n log \left\{b - a\right\}. \tag{1.10}$$

Note que $G(a, b)$ está definida somente para valores de $a < Z_{(1)}$ e $b > Z_{(n)}$. Para resumir, o problema de otimização tri-dimensional da maximização da verossimilhança $L(\underline{Z}|a, m, b)$ reduz-se a um caso bi-dimensional de maximizar $G(a, b)$ sobre a região $a < Z_{(1)}$ e $b > Z_{(n)}$. Da estrutura da verossimilhança, entretanto, pode-se imediatamente concluir que para todos os valores de m tais que $Z_{(1)} < m < Z_{(n)}$, a verossimilhança $L(\underline{\mathbf{Z}}|a, m, b) \to 0$ (e portanto $\log \left\{L(\underline{\mathbf{Z}}|a, m, b) \to \infty\right\}$) quando $a \uparrow Z_{(1)}$ ou $b \downarrow Z_{(n)}$. Portanto, quando um valor modal pode ser observado nos dados (via, por exemplo, um histograma), pode parecer que os EMV para a e b não sejam as estatísticas de ordem $Z_{(1)}$ e $Z_{(n)}$, respectivamente.

Existem algumas rotinas utilizadas na estimação dos parâmetros da distribuição triangular por máxima verossimilhança. Kotz e Dorp (2004) propõe o uso das rotinas *BSearch* e *ABSearch*, conjuntamente. Na prática as estimativas dos parâmetros são obtidas com uso softwares. O software utilizado neste trabalho foi o *MLE Estmator* disponível no sítio:

http://www.seas.gwu.edu/ dorpjr/tab4/publications_book.html.

Um teste da Razão de Verossimilhança

Definição A.1 (Casella e Berger, 2002)**.** *Seja $X_1, ..., X_n$ uma amostra aleatória de uma população com distribuição de densidade de probabilidade $f(x|\theta)$ (θ pode ser um vetor), a*

função de verossimilhança é defina por:

$$L(\theta|x_1,...,x_n) = L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Seja Θ o espaço paramétrico de entrada. A estatística do teste da razão de verossimilhança para testar $H_0 : \theta \in \Theta_0$ contra $H_1 : \theta \in \Theta_0^c$ é

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})}. \tag{1.11}$$

Um teste de razão de verossimilhança (TRV) é algum teste que tem uma região de rejeição da forma $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c \leq 1\}$.

Teste para o caso em que as distribuições são simétricas em torno de 0

Seja $X_1, ..., X_{n_x}$ uma amostra aleatória de uma população com distribuição triangular em $[-a_x, a_x]$ e $Y_1, ..., Y_{n_y}$ uma amostra aleatória de uma população com distribuição triangular em $[-a_y, a_y]$. Ambas com moda igual a *zero*. Como o objetivo é testar a homogeneidade entre as duas amostras e se elas vêm de uma triangular em [-1,1] (isso porque o ganho relativo está definido entre -1 e 1), as hipóteses de interesse podem ser descritas da seguinte maneira:

$$\begin{aligned} H_0 &: a_x = a_y = 1 \\ H_a &: a_x \neq a_y. \end{aligned}$$

Aplicando-se a definição do TRV dada acima, tem-se o seguinte

$$\begin{aligned} \lambda(\mathbf{x}, \mathbf{y}) &= \frac{\max_{a_x = a_y = 1} L(a_x, a_y|\mathbf{x}, \mathbf{y})}{\max_{0 < a_x \neq a_y < 1} L(a_x, a_y|\mathbf{x}, \mathbf{y})} \\ &= \frac{\prod_{i=1}^{r} (x_{(i)} + 1) \prod_{i=r+1}^{n_x} (1 - x_{(i)}) \prod_{j=1}^{s} (y_{(j)} + 1) \prod_{j=r+1}^{n_y} (1 - y_{(j)})}{\frac{1}{\hat{a}_x} \left\{ \prod_{i=1}^{r} \frac{x_{(i)} + \hat{a}_x}{\hat{a}_x} \prod_{i=r+1}^{n_x} \frac{\hat{a}_x - x_{(i)}}{\hat{a}_x} \right\} \frac{1}{\hat{a}_y} \left\{ \prod_{j=1}^{s} \frac{y_{(j)} + \hat{a}_y}{\hat{a}_y} \prod_{j=s+1}^{n_y} \frac{\hat{a}_y - y_{(j)}}{\hat{a}_y} \right\}}. \end{aligned}$$

Apêndice B

Rotina em R para a Metodologia 1

```
#função que calcula o SQE
   calc.sqe <- function(dados){
   # definindo algumas variáveis auxiliares
   ntotal <- length(dados[,1]);
   aux2 <- 0;
   aux3 <- 1;
   aux4 <- 0;
   n <- c(1:2);
   #Calculando o tamanho de cada grupo
   while(aux2 < ntotal){
   n[aux3] <- length(which(dados[,1]==aux4));
   aux2 <- aux2 + n[aux3];
   aux3 <- aux3 + 1;
   aux4 <- aux4 + 1;}
   #calculando a soma do ganho em cada grupo
   tam <- length(n)
   somaganho <- c(1:tam);
   for(i in 1:tam){
   somaganho[i] <- sum(dados[which(dados[,1]==(i-1)),2]);}
```

```
#calculando a soma do ganho ao quadrado em cada grupo
somaganho2 <- c(1:tam);
for (i in 1:tam){
somaganho2[i] <- sum(dados[which(dados[,1]==(i-1)),2]^2);}
#Calculando a variabilidade dentro de cada grupo
Q <- c(1:tam);
for (i in 1:tam){
Q[i] <- (2/(n[i]*(n[i]-1)))*(n[i]*somaganho2[i]-somaganho[i]^2);}
#Calculando a variabilidade entre os grupos e o SQE;
SQE <- 0
C <- matrix(nrow=tam,ncol=tam,0);
for (i in 1:(tam-1)){
for (j in (i+1):tam){
C[i,j] <- (1/(n[i]*n[j]))*(n[j]*(somaganho2[i])+n[i]*(somaganho2[j])
                                -2*somaganho[i]*somaganho2[j]);
SQE <- SQE + n[i]*n[j]*(2*C[i,j] - Q[i] - Q[j]);}}
SQE <- (1/(ntotal*(ntotal-1)))*SQE;
SQE}
# função que faz a re-amostragem por jackknife
calcsqe.jack <- function(dados){
n <- length(dados[,1]);
vetor <- c(1:2)
for (i in 1:n){vetor[i] <- calc.sqe(dados[-i,])}
vetor}
```

Apêndice C

Rotina em R para a Metodologia 2

```
#Calulando a matriz com as comparações
   comparacoes <- function(dados){
   naluno <- max(dados[,1]); #número total de alunos;
   disc <- max(dados[,2]); #número total de disciplinas;
   #número de disc.  que o aluno j foi melhor que o i
   #(posto de i < posto j);
   soma <- matrix(ncol=naluno,nrow=naluno,0);
   #número de disciplinas em comum entre os alunos;
   K <- matrix(ncol=naluno,nrow=naluno,0);
   for(l in 1:disc){
   posto <- c(0); #posição dos alunos que fizeram a disciplina i;
   ndisc <- c(0); #número de vezes que o aluno fez a disciplina i;
   nota <- c(0); #nota obtida pelo aluno na disciplina i;
   posto <- sort(dados[which(dados[,2]==l),1]);
   ndisc <- dados[which(dados[,2]==l),3];
   nota <- dados[which(dados[,2]==l),4];
   tam <- length(posto); # número de alunos que fizeram a disciplina i;
   if(tam>1){
   for(i in 1:(tam-1)){ for(j in (i+1):tam){
```

```
K[posto[i],posto[j]] <- K[posto[i],posto[j]] + 1;
if(ndisc[i]==ndisc[j]){if(nota[j]>nota[i])
{soma[posto[i],posto[j]]<-soma[posto[i],posto[j]]+1; }}
else{if(ndisc[i]>ndisc[j])
{soma[posto[i],posto[j]]<-soma[posto[i],posto[j]]+1;}}
}} }}
RESUL<- soma/K;
RESUL[which(RESUL=="NaN")] <- 0;
RESUL;}
# Calculando grupos
grupos.calc <- function(C2){
num <- max(C2[,1]); grupo <- c(0);
for(i in 1:num){
aux1 <- c(0); aux1 <- C2[which(C2[,1]==i),2];
grupo[i]=aux1[1];}
group <- matrix(nrow=num,ncol=(num+1),0);
for(i in 1:(num-1)){ for(j in (i+1):num){
group[i,j] <- grupo[i]*10 + grupo[j];}}
group[,num+1] <- grupo;
group}
# Calculando SQE
SQE.calc <- function(RESUL,grupo){
lim <- length(grupo[1,]);
max <- max(grupo[,lim]);
n <- c(1:2);
for(i in 1:max){ n[i] <- length(grupo[which(grupo[,lim]==i),lim]);}
ntotal <- sum(n);
tam <- length(n); #número de grupos
Q <- c(1:2);
Cij <- matrix(nrow=tam,ncol=tam,0);
```

```
Cji <- matrix(nrow=tam,ncol=tam,0);
for(l in 1:tam){
aux <- l*10 + l;
Q[l] <- sum(RESUL[which(grupo==aux)]);}
Q <- (2/(n*(n-1)))*Q;
n12 <- 0; n21 <- 0;
for(i in 1:(tam-1)){ for(j in (i+1):tam){
aux1 <- i*10 + j; aux2 <- j*10 + i;
Cij[i,j] <- sum(RESUL[which(grupo==aux1)]);
n12 <- n12 + length(RESUL[which(grupo==aux1)]);
Cji[i,j] <- sum(RESUL[which(grupo==aux2)]);
n21 <- n21 + length(RESUL[which(grupo==aux2)]);}}
for (i in 1:(max-1)){ for(j in (i+1):max){
Cij[i,j] <- (1/(n[i]*n[j]))*Cij[i,j];
Cji[i,j] <- (1/(n[i]*n[j]))*Cji[i,j]; }}
SQE1 <- 0; SQE2 <- 0;
for (i in 1:(tam-1)){ for (j in (i+1):tam){
SQE1 <- SQE1 + n[i]*n[j]*(2*Cij[i,j] - Q[i] - Q[j]);
SQE2 <- SQE2 + n[i]*n[j]*(2*Cji[i,j] - Q[i] - Q[j]); }}
SQE1 <- (1/(ntotal*(ntotal-1)))*SQE1;
SQE2 <- (1/(ntotal*(ntotal-1)))*SQE2;
SQE <- c(SQE1,SQE2);
SQE; }
# Jacknife
jack <- function(RESUL,grupo){
num <- length(RESUL[,1]);
SQE <- matrix(nrow=num,ncol=2,0);
for(i in 1:num){ SQE[i,] <- SQE.calc(RESUL[-i,-i],grupo[-i,-i])}
SQE}
```

Referências Bibliográficas

[1] Agresti, A., and Agresti, B.F.. *Statistical analysis of qualitative variation.* Social Methodology (K.F. Schussler, ed.), 204-237, 1978.

[2] Atkinson, A.B.. *On tehe Measures of Inequality.* journal of Economif Theory, **2**, 244-263, 1970.

[3] Bourguignom, F.. *Decomposable income inequality measures.* Econometrica, **47**, 901-920, 1979.

[4] Bowen, W. and Bok, D.. *The shape of the river: long-term consequences of considering race in Colleg and University admissions.* Princeton, NJ: Princeton University Press, 1998.

[5] Cavalli-Sforza, L.L.. *human diversity.* Proc. XII International Congress of Genetics, Tokyo, **3**, 405-416, 1969.

[6] Chakraborty, R., and Rao, C.R.. *Measurement of genetic variation for evolutionary studies.* Handbook of Statistics **8**, 1991.

[7] Costa, S.. *A construção sociológica da raça no Brasil.* Estudo afro-asiático, **24**(1), 35-61, 2002.

[8] Dachs, J.N.W. and Maia, R.P.. *Subsídios quantitativos para repensar as políticas de acesso à universidade: Aumentando a eqüidade racial e econômica no ensino do terceiro grau do Brasil e no Estado de São Paulo. Primeira parte: Descrição dos alunos da Unicamp que ingressaram no anos de 1994, 1995, 1996 e 1997.* Núcleo de Estudos de Politícas Públicas, Universidade Estadual de Campinas, Relatório Técnico, 2006.

[9] Dachs, J.N.W. and Maia, R.P.. *Subsídios quantitativos para repensar as políticas de acesso à universidade: Aumentando a eqüidade racial e econômica no ensino do terceiro grau do Brasil e no Estado de São Paulo. Segunda parte: Desempenho relativo dos alunos da Unicamp que ingressaram nos anos de 1994, 1995, 1996 e 1997 e descrição dos alunos que prestaram o Provão em 2001.* Núcleo de Estudos de Politícas Públicas, Universidade Estadual de Campinas, Relatório Técnico, 2006.

[10] Dachs, J.N.W. and Maia, R.P.. *Subsídios quantitativos para repensar as políticas de acesso à universidade: Aumentando a eqüidade racial e econômica no ensino do terceiro grau do Brasil e no Estado de São Paulo. Terceira parte: Modelo preditivo para a probabilidade de que um/a jovem brasileiro/a chegue ao ensino superior usando dados da PNAD 1996.* Núcleo de Estudos de Politícas Públicas, Universidade Estadual de Campinas, Relatório Técnico, 2006.

[11] Dagum, C.. *Analysis of income distribution and inequality by education and sex in Canada.* in Advances in Econometrics, **4**, R.L. Basmann and G.F. Rhodes, Jr., Greenwich, CT: JAI Press, 167-227, 1985.

[12] Davison, A.C. and Hinkley, D.V.. *Bootstrap methods and their application*, Cambridge University Press, 1999.

[13] Foster, J.E. and Shneyerov, A.A.. *A general class of additively decomposable inequality measures.* Economic Theory, **44**, 89-111, 1999.

[14] Gini, C.W.. *Variabilita e nutabilita.* Studi Economico-Giuridici della R. Universita di Cogliati **3***(2)*, 3-159, 1912.

[15] Halmos, P.R.. *The theory of unbiased estimation.* Annals of Mathematical Statistics, **17**, 34-43, 1946.

[16] Hoeffding, W.. *A class of statistics with asymptotically normal distribution.* Annals of Mathematical Statistics, **19**, 293-325, 1948.

[17] James, B.J.. *Probabilidade: um Curso em Nível Intermediário.* (Projeto Euclides). Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Segunda Edição, 2002.

[18] Karlin, S., Kennett, R., and Bonne-Tamir, B.. *Analysis of biochemical genetic data on Jewish populations: II. Results and interpretations of heterogeneity indices and distance measures with respect to standards.* American journal of Human Genetics, **31**, 341-365, 1979.

[19] Kotz, S. and Dorp, J.R. van. *Beyond Beta, Other Continuous Families of Distributions with Bounded Support and Applications*, World Scientific Press, Singapore, 2004.

[20] Lee, A.J.. *U-Statistics - Theory and Pratice.* Marcel Dekker, Nova Iorque, NY, 1990.

[21] Leite, J.G. and Singer, J.M.. *Métodos Assintóticos em Estatísticas - Fundamentos e Aplicações.* AAssociação Brasileira de Estatística, 9° Simpósio Nacional de Probabilidade e Estatística, São Paulo, 1990.

[22] Lehmann, E.L.. *Robust estimation in Analysis of Variation.* Annals of Mathematical Statistics, **34**, 957-966, 1963.

[23] Lehmann, E.L.. *Elements of Large-Sample Theory.* Springer-Verlag, Nova Iorque, NY, 1999.

[24] Mahalanobis, P.. *On the generalized distance in statistics.* Proceedings of the National Institute of Sciences of India, **2**, 49-55, 1936.

[25] Nayak, T.K.. *An analysis of diversity using Rao's quadratic entropy.* Sankya B, **48**, 315-330, 1986.

[26] Nayak, T.K., and Gastwirth, J.L.. *The use of diversity analysis to asses the relative influence factrs affecting the income distributin.* Journal of Business & Economic Statistics, **7**(4), 453-460, 1989.

[27] Nei, M.. *Estimation of average heterozygosity and genetic distance from small number of individuals.* Genetics, **89**, 583-590, 1978.

[28] Patil, G.P. and Taillie, C.. *Diversity as a concept and its measurement.* Journal of the American Statistical Association, **77**(379), 548-561, 1982.

[29] Pedrosa, R.H.L., Dachs, J.N.W., Maia, R.P., Andrade, C.Y., Carvalho, S.C.. *Academic Performance, Students' Background and Affirmative Action at a Brazilian Research University*, Higher Education Management and Policy, Vol.19, Issue 3, 2007.

[30] Pielou, E.C.. *Ecological Diversity*. Wiley & Sons, Nova Iorque, 1975.

[31] Pinheiro, H.P., Seiller-Moiseiwitsh, F., and Sen, P.K.. *Analysis of variance for Hamming distances applied to unbalanced designs*. Research Report No.30/01, Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Brazil, 2001.

[32] Pinheiro, H.P., Pinheiro, A., and Sen, P.K.. *Comparison of genomic sequences using Hamming distance*. Journal of Statistical Planning and Inference, **130**(1-2), 325-339, 2005.

[33] Pinheiro, A., Pinheiro, H.P., and Sen, P.K.. *The use Hamming distance in bioinformatics*. Handbook of Statistics: Bioinformatics. (no prelo), 2008.

[34] Pinheiro, A., Sen, P.K., and Pinheiro, H.P.. *Parametric modeling of genomic sequences distance*. Calcutta Statistical Association Bulletin, **58**(229-230), 1-14, 2006.

[35] Pinheiro, A., Sen, P.K., and Pinheiro, H.P.. *Decomposability og high-Dimensional Diversity Measures: Quasi U-Statistics, Martingales and Nonstandard Asymptotics*. Relatório de Pesquisa, IMECC/UNICAMP, Agosto, 2006.

[36] Pinheiro, A. and Pinheiro H.P.. *Métodos Estatísticos Não-Paramétricos e suas Aplicações*. 26° Colóquio Brasileiro de Matemática, Publicações Matemáticas, Instituto Nacional de Matemática Pura e Aplicada, Rio de Janeiro, 2007.

[37] Rao, C.R.. *Diversity: Its measurement, decomposition, apportionment and analysis*. Sankya A, **44**, 1-21, 1982.

[38] Rao, C.R.. *Gini-Simpson index of diversity: A characterization, generalization and applications*. Utilitas Mathematica, **21**, 273-282, 1982.

[39] Rao, C.B.. *Convexity Properties of Entropy functions and analysis of diversity*. Lecture Notes - Monograph Series, **5**, 64-77, 1984.

[40] Randles, R.H. and Wolfe, D.A.. *Introduction to the Theory of Comparametric Statistics.* Krieger Publishing Company, Malabar, Florida, 1991.

[41] Sen, A.. *on Economic Inequality.* Clarendon Press, Oxford, 1973.

[42] Sen, P.K.. *Utility-oriented Simpson-tyoe indexes and inequality measures.* Calcuta Statistical Association Bulletin, **49**, 1-22, 1999.

[43] Sen, P.K., and Singer, J.M.. *Large Sample Methods in Statistics - An Introdution with Applications.* Chapman & Hall, Nova Iorque, 1993.

[44] Shangvi, l.D.. *Comparison of genetical and morphological methods for a study of biological diferences.* American Journal of Physical Antropology, **11**, 385-404, 1953.

[45] Shorrocks, A.F.. *The class of additively decomposable inequality measures.* Econometrica, **48**, 613 - 615, 1980.

[46] Peter, R.R., and Sneath, P.H.A.. *Principles of Numerical Tazonomy.* W.H.Freeman, Nova Iorque, NY, 1963.

[47] Theil, H.. *Economic and Information Theory.* Amsterdam,: North-Holland, 1967.